Multifaceted Neural News Recommendation: Common Components, Bias Mitigation, and Multilinguality

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von
Andreea Iana
aus Bukarest

Mannheim, 2026

Dekan:	Prof. Dr. Claus Hertling, Universität Mannheim
Referent:	Prof. Dr. Heiko Paulheim, Universität Mannheim
Korreferent:	Prof. Dr. Goran Glavaš, Universität Würzburg
Korreferent:	Prof. Dr. Julia Neidhardt, Technische Universität Wien

Tag der mündlichen Prüfung: 30. March 2026

To my loving family

ACKNOWLEDGMENTS

A Ph.D. journey is much like a hike up a beautiful mountain: long and strenuous climbs, unexpected descents, and then the ascent once more; sometimes losing sight of the summit, only to be rewarded with a breathtaking view upon reaching it. While standing at the peak is a tremendous achievement in itself, the journey upward is equally meaningful and far more enjoyable, especially when shared with wonderful companions and enriched by unforgettable moments. For this reason, I would like to express my heartfelt gratitude to everyone who has accompanied, supported, advised, and inspired me during my Ph.D. journey.

First and foremost, I would like to express my wholehearted gratitude to my doctoral supervisor, Prof. Heiko Paulheim. You supported and believed in me from my early days as a student assistant and throughout my doctoral studies. Your academic guidance, moral support whenever needed, and the freedom you gave me to explore and find my research interests were invaluable. I am especially grateful for your openness to my ideas and for being a dedicated supervisor and exceptional mentor: one who listens patiently and offers thoughtful advice. Knowing that I could always turn to you gave me confidence and made the challenges of my Ph.D. journey easier to navigate.

I would like to extend my sincere gratitude to my co-supervisor, Prof. Goran Glavaš. You stepped in with insightful advice and support when I needed it most, and you consistently encouraged me while sharing inspiring research ideas. I have learned a great deal from your approach to research and academic writing. I am deeply grateful for your exceptional dedication as a supervisor, as well as your generous mentorship and the many engaging stories you shared throughout our collaboration. In addition, I would like to thank my third committee member, Prof. Julia Neidhardt, for the time and involvement in reviewing my dissertation.

I would also like to thank all my collaborators throughout the years, particularly Dr. Mehwish Alam, Katharina Ludwig, and Dr. Philipp Müller, for the fruitful discussions and enjoyable collaborations, which were particularly valuable during the early stages of my Ph.D.

My Ph.D. journey has also been significantly shaped by the people around me. I am fortunate to have been surrounded by such wonderful colleagues and friends: Adrian, Rita, Martin, Franz, Julie, Sven, and Ezgi. Thank you for the many conversations and moments of laughter during the tea or coffee breaks, and all the other memorable times we shared. I am especially grateful to my dear friends, Eduardo and Gio, for always cheering me on and asking about my deadlines.

În încheiere, aș dori să le mulțumesc părinților mei, care mi-au inspirat pasiunea pentru cercetare si au fost întotdeauna alături de mine, susținându-mă la fiecare pas, fără ezitare. Nimic din ceea ce am realizat până acum nu ar fi fost posibil fără dragostea, răbdarea, încurajarea și sacrificiile voastre. Această reușită este la fel de mult a voastră, cât și a mea.

Last but not least, words cannot express how much I want to thank you, Fabian, for your love, patience, unwavering moral support and advice, and being by my side throughout all these years.

Abstract

The consumption of news has shifted significantly over the past decade from traditional outlets, such as print newspapers and television, to digital platforms, including news websites, aggregators, and social media. This transformation has made personalized recommender systems central to how users access information, as the overwhelming volume and velocity of online content render manual discovery of relevant articles impractical. These systems learn from users' reading behaviors to filter and curate content that matches their interests. Neural approaches are particularly well-suited to this domain, as they can model the rich textual content of news, capture dynamic user interests, and learn complex behavioral patterns from large-scale interaction data.

While personalized neural news recommenders (NNRs) enhance engagement and help users efficiently discover relevant content, they also act as powerful gatekeepers that shape users' exposure to information. When optimized solely for relevance or engagement, they can inadvertently narrow individuals' information diets, reinforce existing biases, and reduce exposure to diverse viewpoints. At the same time, the growing linguistic diversity of the Web exposes a critical inequity. The dominance of high-resource languages such as English in the training data, benchmark datasets, and model development, combined with the field's near-exclusive focus on monolingual news consumption, results in systems that inadequately serve both speakers of low-resource languages (e.g., languages with small digital footprints) and multilingual users whose cross-lingual reading behaviors are not well supported by predominantly monolingual recommenders. Addressing these intertwined technical and societal challenges requires a multifaceted approach: one that moves beyond relevance-only optimization to integrate responsible curation that mitigates biases and fosters diverse information exposure, alongside inclusive multilingual design to ensure equitable access across languages.

Motivated by the preceding considerations, this thesis presents new analyses, methods, and resources aimed at advancing neural news recommender systems that are more efficient, responsible, and inclusive. Concretely, we segment our contributions into three areas.

(1) Analysis of Common Components. We first conduct two comprehensive analyses that decouple and evaluate the common components of NNRs to understand how their design choices and interactions affect downstream performance. We introduce a unified framework for systematically comparing NNRs across three key dimensions: user modeling approaches, click behavior fusion strategies, and training objectives. Our proposed late fusion method, based on the mean-pooling of dot-product scores between candidate and clicked news, outperforms more complex user encoders while greatly reducing model complexity. Building on these insights, we develop a multifaceted evaluation protocol to examine news and user encoder architectures in terms of representational similarity, recommendation overlap, and overall recommendation accuracy. Our findings challenge the assumption that architectural complexity is essential for strong performance, demonstrating instead that accuracy depends primarily on the semantic quality of news representations, while user encoders can be significantly simplified without loss of effectiveness.

(2) Multi-aspect Algorithmic Bias Analysis and Mitigation. The second part examines the societal dimension of recommendation. Specifically, we analyze how news recommenders introduce or amplify algorithmic biases and propose a modular framework to mitigate them through multi-aspect customization. We first investigate the sentiment and stance biases using a self-compiled German news corpus on migration, finding that recommendation models exhibit a slight tendency to suggest negatively framed articles and to align with users' existing preferences, revealing a risk of reinforcing opinions and reducing content diversity. Extending this analysis to bilingual German-English corpora and to more complex neural models, we further identify a consistent trade-off between personalization and diversity, as well as stronger alignment between topical and political personalization than between sentiment and content relevance. However, the majority of beyond-relevance optimization methods typically target only a single dimension – either personalization or diversification – "hardcoded" into the model architecture or training objective, which limits flexibility when multiple user criteria

need to be satisfied simultaneously. To address this rigidity, we introduce MANNeR, a modular multi-aspect neural news recommendation framework that leverages metric-based learning to train separate news encoders for distinct aspects such as content, topic, or sentiment. By linearly combining aspect-specific similarity scores at ranking time, our model enables dynamic control over content relevance and aspect-based diversity trade-offs, and supports ad-hoc customization without retraining. We demonstrate that MANNeR outperforms existing NNRs across content-based recommendation, as well as single- and multi-aspect diversification and personalization tasks. We further show that when equipped with a multilingual pretrained language model backbone, its news encoder remains robust in cross-lingual transfer, namely when the model trained on a source language is directly applied to the target language without additional fine-tuning.

(3) Multilinguality in News Recommendation. The final part addresses linguistic inclusivity. Language fundamentally shapes both access to information, and the nature of user engagement, yet existing news recommendation datasets are predominantly monolingual or support only a handful of high-resource languages. To foster more inclusive news recommendation technology, we introduce xMIND, the first multilingual, multi-parallel news recommendation benchmark, derived via machine translation from the widely used English MIND dataset. Spanning 14 linguistically and geographically diverse languages, xMIND enables direct performance comparison across high- and low-resource languages and supports the simulation of multilingual news consumption. We systematically evaluate several NNRs under zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer scenarios – where the NNRs are tested on new languages without (ZS-XLT) or with limited (FS-XLT) fine-tuning on a few target-language instances. Our results reveal substantial performance degradation in resource-lean and typologically distant target languages, and limited benefits from random or topic-based target-language data injection during training. To tackle these limitations, we domain-specialize a pretrained multilingual sentence encoder using denoising auto-encoding and machine translation objectives on self-compiled multilingual news corpora. The resulting news-adapted sentence encoder (NaSE) improves cross-lingual transfer performance when used as the NNR backbone, particularly in ZS-XLT settings, and remains effective even when frozen during downstream fine-tuning. These findings challenge the established paradigm of fine-tuning the NNR's underlying language model on user click data. Finally, we present a lightweight and strong baseline that combines frozen NaSE embeddings with late click behavior fusion, achieving state-of-the-art performance in true cold-start and few-shot multilingual news recommendation.

We are confident that our contributions to neural news recommendation will catalyze future work towards the development of efficient, responsible, and inclusive recommendation technology that equitably serves users across languages and contexts.

Zusammenfassung

Die Nachrichtenkonsum hat sich im vergangenen Jahrzehnt deutlich von traditionellen Medien wie Zeitungen und Fernsehen hin zu digitalen Plattformen wie Nachrichtenwebseiten, Aggregatoren und sozialen Medien verlagert. Diese Transformation hat personalisierte Empfehlungssysteme (Recommender Systems, RS) zu einem zentralen Bestandteil des Informationszugangs gemacht, da die überwältigende Menge und Geschwindigkeit von Online-Inhalten die manuelle Entdeckung relevanter Artikel unpraktikabel machen. Solche Systeme lernen aus dem Leseverhalten der Nutzer*innen, um Inhalte zu filtern und zu kuratieren, die ihren Interessen entsprechen. Neuronale Netze eignen sich besonders gut für diese Bereich, da sie den reichhaltigen Textinhalt von Nachrichten modellieren, dynamische Nutzerinteressen erfassen und komplexe Verhaltensmuster aus großen Interaktionsdaten lernen können.

Während personalisierte neuronale Nachrichtenempfehlungssysteme (Neural News Recommender Systems, NNRs) das Engagement steigern und Nutzern helfen, relevante Inhalte effizient zu entdecken, fungieren sie gleichzeitig als mächtige Gatekeeper, die bestimmen, welchen Informationen Nutzer*innen ausgesetzt sind. Wenn sie ausschließlich auf Relevanz oder Engagement optimiert werden, können sie ungewollt die Informationsdiät des Personen verengen, bestehende Biases verstärken und die Exposition gegenüber diversen Standpunkten verringern. Gleichzeitig offenbart die wachsende sprachliche Vielfalt des Webs eine kritische Ungleichheit. Die Dominanz von ressourcenreichen Sprachen wie Englisch in den Trainingsdaten, Benchmark-Datensätzen und der Modellentwicklung, kombiniert mit einem fast ausschließlichen Fokus des Fachgebiets auf einsprachigen Nachrichtenkonsum, führt zu Systemen, die sowohl Sprecher*innen von Sprachen mit geringen digitalen Ressourcen als auch mehrsprachige Nutzer*innen, deren sprachübergreifendes Leseverhalten von vorwiegend monolingualen Empfehlungssystemen unzureichend unterstützt wird, unzureichend bedienen. Die Bewältigung dieser verflochtenen technischen und gesellschaftlichen Herausforderungen erfordert einen facettenreichen Ansatz: einen, der über reine Relevanzoptimierung hinausgeht, um eine verantwortungsvolle Kuratierung zu integrieren, die Biases mindert und vielfältige Nachrichtenauswahl fördert, sowie ein inklusives, mehrsprachiges Design, um einen gleichberechtigten Zugang über Sprachen zu gewährleisten.

Ausgehend von diesen vorangegangenen Überlegungen präsentiert diese Doktorarbeit neue Analysen, Methoden und Ressourcen zur Weiterentwicklung effizienter, verantwortungsvoller und inklusiver neuronaler Nachrichtenempfehlungssysteme. Konkret gliedern wir unsere Beiträge in drei Bereiche.

(1) Analyse gemeinsamer Komponenten. Zunächst führen wir zwei umfassende Analysen durch, die die gemeinsamen Komponenten von NNRs entkoppeln und evaluieren, um zu verstehen, wie Designentscheidungen und deren Zusammenspiel die Leistung beeinflussen. Wir stellen ein einheitliches Framework vor, das NNRs systematisch entlang dreier Schlüsseldimensionen vergleicht: Ansätze zur Benutzermodellierung, Strategien zur Fusion des Klickverhaltens und Trainingsziele. Unsere vorgeschlagene *Late Fusion*-Methode, basierend auf *mean-pooling* der Skalarproduktwerte zwischen Kandidaten- und geklickten Nachrichten, übertrifft komplexere User Encoders, während sie die Modellkomplexität erheblich reduziert. Aufbauend auf diesen Erkenntnissen entwickeln wir ein mehrdimensionales Evaluationsprotokoll zur Untersuchung von Nachrichten- und Benutzerencoder-Architekturen hinsichtlich Repräsentationsähnlichkeit, Überlappung der Empfehlungen und Gesamtgenauigkeit der Empfehlungen. Unsere Ergebnisse stellen die Annahme in Frage, dass architektonische Komplexität für eine starke Leistung essenziell ist. Stattdessen zeigen wir, dass die Genauigkeit primär von der semantischen Qualität der Nachrichtenrepräsentationen abhängt, während User Encoders ohne Effektivitätsverlust deutlich vereinfacht werden können.

(2) Multiaspektuelle algorithmische Bias-Analyse und -Minderung. Der zweite Teil untersucht die gesellschaftliche Dimension von Empfehlungssystemen. Insbesondere analysieren wir, wie Nachrichtenempfehlungssysteme algorithmische Biases einführen oder verstärken, und schlagen ein modulares

Framework vor, um diese durch Anpassung hinsichtlich unterschiedlicher Aspekte zu mindern. Wir untersuchen zunächst Sentiment- und Standpunkt-Biases anhand eines selbst zusammengestellten deutschen Nachrichtenkorpus zum Thema Migration. Dabei zeigen wir, dass Empfehlungsmodelle eine leichte Tendenz zeigen, negativ formulierte Artikel vorzuschlagen, und sich an den bestehenden Nutzerpräferenzen auszurichten – was das Risiko von Meinungsverstärkung und geringerer Inhaltsvielfalt erhöht. Durch die Erweiterung der Analyse auf ein deutsch-englisches Korpus und komplexere neuronale Modelle identifizieren wir zudem einen konsistenten Trade-off zwischen Personalisierung und Vielfalt sowie eine stärkere Ausrichtung zwischen thematischer und politischer Personalisierung als zwischen Sentiment und Inhaltsrelevanz. Die meisten Optimierungsmethoden, die über die Relevanz hinausgehen, zielen jedoch typischerweise nur auf eine einzige Dimension ab – entweder Personalisierung oder Diversifizierung – die fest in der Modellarchitektur oder das Trainingsziel verankert ist und somit wenig Flexibilität bietet, wenn mehrere Nutzerkriterien gleichzeitig erfüllt werden müssen. Um diese Starrheit zu überwinden, stellen wir MANNeR vor, ein modulares multiaspektuelles neuronales Nachrichtenempfehlungssystem, das metrikbasiertes Lernen nutzt, um separate Nachrichtenencoder für verschiedene Aspekte wie Inhalt, Thema oder Sentiment zu trainieren. Durch lineare Kombination aspekt-spezifischer Ähnlichkeitswerte zum Zeitpunkt des *Rankings* ermöglicht unser Modell dynamische Kontrolle über *Trade-offs* zwischen Inhaltsrelevanz und aspektbasierter Vielfalt und unterstützt ad-hoc-Anpassungen ohne erneutes Training. Wir zeigen, dass MANNeR bestehende NNRs sowohl bei inhaltsbasierten Empfehlungen als auch bei Ein- oder Mehraspekt-Diversifizierung und Personalisierungsaufgaben übertrifft. Weiterhin zeigen wir, dass die Nachrichtenencoder, wenn er mit einem multilingualen vortrainierten Sprachmodell (Pretrained Language Model, PLM) ausgestattet ist, robust in sprachübergreifendem Transfer (Cross-Lingual Transfer, XLT) bleibt, d.h. wenn das auf einer Quellsprache trainierte Modell ohne zusätzliches *Fine-tuning* direkt auf die Zielsprache angewendet wird.

(3) Mehrsprachigkeit in der Nachrichtenempfehlung. Der letzte Teil befasst sich mit der sprachlicher Inklusivität. Sprache beeinflusst grundlegend sowohl den Zugang zu Informationen als auch die Art der Nutzerinteraktion, doch existierende Datensätze für Nachrichtenempfehlung sind überwiegend einsprachig oder unterstützen nur wenige ressourcenreichen Sprachen. Um inklusivere Nachrichtenempfehlungstechnologie zu fördern, führen wir xMIND ein – den ersten mehrsprachigen, multiparallelen Benchmark für Nachrichtenempfehlung, der mittels maschineller Übersetzung aus dem weit verbreiteten englischen MIND-Datensatz abgeleitet wurde. xMIND umfasst 14 sprachlich und geografisch vielfältige Sprachen und ermöglicht einen direkten Leistungsvergleich von NNRs für ressourcenreichen und ressourcenarmen Sprachen sowie die Simulation multilingualem Nachrichtenkonsum. Wir evaluieren mehrere NNRs systematisch in *zero-shot* (ZS-XLT) und *few-shot* (FS-XLT) Szenarien des Sprachübergreifenden Transfers, bei denen die NNRs auf neuen Sprachen ohne (ZS-XLT) oder mit begrenztem (FS-XLT) *Fine-tuning* auf einigen wenigen zielsprachlichen Trainingsinstanzen getestet werden. Unsere Ergebnisse offenbaren eine erhebliche Leistungsverschlechterung in ressourcenarmen und typologisch weit entfernten Zielsprachen sowie begrenzte Vorteile durch zufällige oder themenbasierte zielsprachliche Dateninjektion während des Trainings. Um diese Einschränkungen zu überwinden, spezialisieren wir einen vortrainierten multilingualen Satzencoder für den Nachrichtenbereich mittels *denoising Auto-encoding* und maschinellen Übersetzungszielen auf selbst zusammengestellten multilingualen Nachrichtencorpora. Der resultierende Nachrichten-angepasster Satzencoder (NaSE) verbessert die Sprachübergreifende Transferleistung, wenn er als NNR-Backbone verwendet wird, insbesondere in ZS-XLT-Szenarien, und bleibt auch dann effektiv, wenn er während des nachfolgenden *Fine-tuning* eingefroren (*frozen*) wird. Diese Ergebnisse stellen das etablierte Paradigma in Frage, das zugrundeliegende Sprachmodell des NNR auf Nutzer-Klickdaten zu fine-tune. Abschließend präsentieren wir eine ressourcenschonende und leistungsstarke Baseline, die *frozen* NaSE-Embeddings mit *late click behavior fusion* kombiniert und damit Bestleistungen in echten *Cold-Start*- und *Few-Shot*-Szenarien der multilingualen Nachrichtenempfehlung erzielt.

Wir sind zuversichtlich, dass unsere Beiträge zur neuronale Nachrichtenempfehlungstechnologie zukünftige Arbeiten zur Entwicklung effizienter, verantwortungsvoller und inklusiver Empfehlungstechnologien anregen, die Nutzer*innen über Sprachen und Kontexte hinweg gerecht dienen.

CONTENTS

List of Publications xvii

List of Figures xxii

List of Tables xxiv

List of Acronyms xxv

1 Introduction 1
- 1.1 Motivation . . . 1
  - 1.1.1 Personalized Neural News Recommendation . . . 3
  - 1.1.2 The Case for Multifaceted Neural News Recommendation . . . 5
  - 1.1.3 Challenges . . . 10
- 1.2 Contributions . . . 11
- 1.3 Outline . . . 13

I FUNDAMENTALS 15

2 Representation Learning 17
- 2.1 Language Modeling . . . 17
  - 2.1.1 From Statistical to Neural Language Models . . . 18
  - 2.1.2 Static Word Embedding Models . . . 19
  - 2.1.3 Contextualized Representations . . . 20
- 2.2 Pretrained Language Models . . . 25
  - 2.2.1 Language Model Pretraining . . . 25
  - 2.2.2 Sequence-Level Representations . . . 27
- 2.3 Multilingual Representation Learning . . . 27
  - 2.3.1 Multilingual Pretraining . . . 27
  - 2.3.2 Multilingual Sentence Encoders . . . 28

3 Neural News Recommendation 31
- 3.1 Personalized Neural News Recommendation . . . 31
  - 3.1.1 News Modeling . . . 32
  - 3.1.2 User Modeling . . . 37
  - 3.1.3 Personalized Ranking . . . 42
  - 3.1.4 Model Training . . . 42
  - 3.1.5 Selected Baselines . . . 44
- 3.2 Evaluation Protocol . . . 46
  - 3.2.1 Evaluation Metrics . . . 46
  - 3.2.2 Benchmark Datasets . . . 48

3.3 Evaluation Frameworks . . . 51
3.3.1 Open-Source Libraries . . . 51
3.3.2 NewsRecLib . . . 53
3.4 Responsible News Recommendation . . . 57
3.4.1 Debiasing . . . 58
3.4.2 Diversity . . . 59
3.5 Multilinguality . . . 60

II ANALYSIS OF COMMON COMPONENTS 65

4 User Modeling and Training Objectives 67
4.1 Introduction . . . 67
4.2 Methodology . . . 69
4.2.1 User Modeling . . . 69
4.2.2 Click Behavior Fusion . . . 70
4.2.3 Training Objectives . . . 71
4.3 Experimental Setup . . . 71
4.4 Results and Discussion . . . 71
4.5 Conclusion . . . 74

5 Encoder Selection 75
5.1 Introduction . . . 75
5.2 Related Work . . . 76
5.3 Methodology . . . 77
5.3.1 Encoders of Neural News Recommenders . . . 77
5.3.2 Similarity Evaluation . . . 79
5.4 Experimental Setup . . . 80
5.5 Results and Discussion . . . 81
5.5.1 News Encoder Architectures . . . 81
5.5.2 User Encoder Architectures . . . 83
5.5.3 Key Takeaways . . . 85
5.6 Conclusion . . . 86

III MULTI-ASPECT BIAS ANALYSIS AND MITIGATION 87

6 Multi-Aspect Bias Analysis in News Recommenders 89
6.1 Introduction . . . 90
6.2 Related Work . . . 91
6.3 GeNeG: German News Dataset . . . 92
6.3.1 News Data Collection and Annotation . . . 92
6.3.2 User Data Collection . . . 94
6.4 Preliminary Bias Analyses . . . 95
6.4.1 Experimental Setup . . . 95
6.4.2 Results and Discussion . . . 96
6.5 NeMig: Bilingual News Dataset . . . 100
6.5.1 News Data Collection and Annotation . . . 100
6.5.2 Knowledge Graph Construction . . . 104
6.5.3 User Data Collection . . . 105
6.5.4 Data Availability . . . 106

6.6 Bias Analyses Revisited . . . 107
6.6.1 Experimental Setup . . . 107
6.6.2 Results and Discussion . . . 109
6.7 Conclusion . . . 111

7 Modular Multi-Aspect Neural News Recommendation 113
7.1 Introduction . . . 114
7.2 Related Work . . . 114
7.3 MANNeR: Multi-Aspect Neural News Recommendation . . . 115
7.3.1 News Encoder . . . 116
7.3.2 Modular Training . . . 116
7.3.3 Inference: Custom Ranking Functions . . . 117
7.4 Experimental Setup . . . 117
7.5 Results and Discussion . . . 119
7.5.1 Content Personalization . . . 119
7.5.2 Single-Aspect Customization . . . 119
7.5.3 Multi-Aspect Customization . . . 122
7.5.4 Cross-Lingual Transfer . . . 123
7.5.5 Computational Complexity . . . 123
7.6 Conclusion . . . 124

IV MULTILINGUALITY 125

8 Cross-Lingual News Recommendation 127
8.1 Introduction . . . 128
8.2 Related Work . . . 129
8.2.1 News Recommendation . . . 129
8.2.2 LLM-based Recommendation . . . 130
8.2.3 Recommendation Datasets . . . 130
8.2.4 Cold-Start Recommendation . . . 131
8.3 xMIND Dataset . . . 131
8.4 Experimental Setup . . . 134
8.4.1 Benchmarked Neural News Recommenders . . . 135
8.4.2 Training and Optimization Details . . . 135
8.4.3 Cross-Lingual Recommendation Scenarios . . . 136
8.5 Zero-Shot Cross-lingual Transfer . . . 137
8.6 Few-Shot Cross-lingual Transfer . . . 140
8.6.1 Random Replacement Replacement of Source-Language News . . . 140
8.6.2 Category-Based Replacement of Source-Language News . . . 142
8.6.3 Random Versus Category-Based Replacement of Source-Language News . . . 143
8.7 Further Analysis of Cross-lingual Transfer in Neural News Recommendation . . . 144
8.7.1 Fine-Tuning Strategies . . . 144
8.7.2 Representation Alignment Across Languages . . . 146
8.8 Dataset Analysis . . . 149
8.8.1 Manual Quality Estimation of Translation . . . 150
8.8.2 Robustness of NNRs to Translation Quality . . . 151
8.9 Conclusion . . . 153

9 News Domain Adaptation of Sentence Encoders **155**
9.1 Introduction . . . 155
9.2 Related Work . . . 156
9.3 Domain-Specialized Corpora . . . 157
9.4 Methods for Domain Specialization . . . 158
9.4.1 Domain Adaptation . . . 158
9.4.2 Training Details . . . 159
9.5 Experimental Setup . . . 160
9.6 Results and Discussion . . . 161
9.7 Conclusion . . . 164

V CONCLUSIONS AND PERSPECTIVES **167**

10 Conclusions and Perspectives **169**
10.1 Summary of Findings . . . 169
10.2 Perspectives on Effective, Responsible, and Inclusive (News) Recommendation . . . 172

Bibliography **177**

A Published Resources **231**

B Wikipedia over Time **233**

C NewsRecLib Logging Details **235**
C.1 Configuration Logging . . . 235
C.2 Model Metadata Logging . . . 239

D Additional Results for Chapter 4 **241**

E Experimental Details for Chapter 7 **243**
E.1 Reproducibility . . . 243
E.1.1 Model Parameters . . . 243
E.1.2 Hyperparameters and Implementation . . . 243
E.2 Additional Results . . . 244
E.2.1 Content Personalization . . . 244
E.2.2 Single-Aspect Customization . . . 245
E.2.3 Multi-Aspect Customization . . . 245
E.2.4 Cross-Lingual Transfer . . . 245
E.2.5 Time Complexity Analysis . . . 245

F Experimental Details for Chapter 8 **247**
F.1 Language Model Support . . . 247
F.2 Hyperparameter Optimization . . . 247

G Experimental Details for Chapter 9 **249**
G.1 PolyNews Corpora . . . 249
G.1.1 Language Characteristics . . . 249
G.1.2 Parallel Corpora Statistics . . . 251
G.2 Reproducibility . . . 251
G.3 Additional Results . . . 252

List of Publications

Most of the work presented in this thesis, including text, figures, and tables has been previously published in the proceedings of top-tier international conferences and workshops. We reference the publications in the subsequent Chapters and list them below in inverse chronological order. The publication venues are annotated according to their CORE Conference Ranking (A*, A), a widely adopted ranking system for conferences in Computer Science. The designation A* indicates a flagship conference within a discipline, and A denotes a highly respected venue in that discipline area.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2025. Multilinguality in MIND: Advancing Cross-Lingual News Recommendation with a Multilingual Dataset. *ACM Transactions on Recommender Systems*.

Andreea Iana, Fabian David Schmidt, Goran Glavaš, and Heiko Paulheim. 2025. News Without Borders: Domain Adaptation of Multilingual Sentence Embeddings for Cross-lingual News Recommendation. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval* (ECIR 2025, A), pages 119–137. Springer Nature Switzerland.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. Peeling Back the Layers: An In-depth Evaluation of Encoder Architectures in Neural News Recommenders. In *INRA 2024: Proceedings of the International Workshop on News Recommendation and Analytics co- located with the 2024 ACM Conference on Recommender Systems (RecSys 2024)*, volume Vol-3929, pages 1–15, Aachen, Germany. RWTH Aachen.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. Train Once, Use Flexibly: A Modular Framework for Multi-Aspect Neural News Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (EMNLP 2024, A*), pages 9555–9571, Miami, Florida, USA. Association for Computational Linguistics.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2024, A*), pages 553–563, New York, NY, USA. Association for Computing Machinery.

Andreea Iana, Mehwish Alam, and Heiko Paulheim. 2024. A Survey on Knowledge-aware News Recommender Systems. *Semantic Web*, 15(1):21–82.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023. NewsRecLib: A PyTorch-Lightning Library for Neural News Recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (EMNLP 2023, A*), pages 296–310, Singapore. Association for Computational Linguistics.

Andreea Iana, Mehwish Alam, Alexander Grote, Nevena Nikolajevic, Katharina Ludwig, Philipp Müller, Christof Weinhardt, and Heiko Paulheim. 2023. NeMig - A Bilingual News Collection and

Knowledge Graph about Migration. In *INRA 2023: Proceedings of the International Workshop on News Recommendation and Analytics co-located with the 2023 ACM Conference on Recommender Systems (RecSys 2023)*, volume 3561, pages 1–22, Aachen, Germany. RWTH Aachen.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023. Simplifying Content-Based Neural News Recommendation: On User Modeling and Training Objectives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2023, A*), pages 2384–2388, New York, NY, USA. Association for Computing Machinery.

Mehwish Alam*, **Andreea Iana***, Alexander Grote, Katharina Ludwig, Philipp Müller, and Heiko Paulheim. 2022. Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection. In *Companion Proceedings of the Web Conference 2022, WWW '22*, page 448–457, New York, NY, USA. Association for Computing Machinery.

*Both authors contributed equally.

The publications listed before are directly related to the content of this dissertation and are used in § 3.1.5, § 3.2.2, Section 3.3, and Chapter 4 to Chapter 9. Moreover, the author contributed to and published other research work during the course of her doctoral studies, which we list here in inverse chronological order.

Andreea Iana, Célina Treuillier, Vandana Yadav, Benjamin Kille, Andreas Lommatzsch, Özlem Özgöbek. 2025. The 13th International Workshop on News Recommendation and Analytics (INRA 2025). In *Proceedings of the Nineteenth ACM Conference on Recommender Systems* (RecSys 2025, A), pages 1399–1403, New York, NY, USA. Association for Computing Machinery.

Martin Böckling, Heiko Paulheim, and **Andreea Iana**. 2025. Walk&Retrieve: Simple Yet Effective Zero-shot Retrieval-Augmented Generation via Knowledge Graph Walks. In *Proceedings of the Workshop Information Retrieval's Role in RAG Systems (IR-RAG 2025) co-located with the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)*.

Daniel Delev, Emmie Schiffer, Felix Vogl, **Andreea Iana**, Heiko Paulheim. 2024. Towards Improving a Student Advisory Service Chatbot Using Knowledge Graphs. In *Proceedings of the International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs (RAGE-KG 2024) co-located with the 23rd International Semantic Web Conference (ISWC 2024)*.

Katharina Ludwig, Alexander Grote, **Andreea Iana**, Mehwish Alam, Heiko Paulheim, Harald Sack, Christof Weinhardt, and Philipp Müller. 2023. Divided by the Algorithm? The (Limited) Effects of Content- and Sentiment-Based News Recommendation on Affective, Ideological, and Perceived Polarization. Social Science Computer Review, 41(6):2188– 2210.

Laura Waltersdorfer, Anna Breit, Fajar J Ekaputra, Marta Sabou, Andreas Ekelhart, **Andreea Iana**, Heiko Paulheim, Jan Portisch, Artem Revenko, Annette ten Teije, et al. 2023. Semantic Web Machine Learning Systems: An Analysis of System Patterns. In Compendium of Neurosymbolic Artificial Intelligence, pages 77–99. IOS Press.

Anna Breit, Laura Waltersdorfer, Fajar J. Ekaputra, Marta Sabou, Andreas Ekelhart, **Andreea Iana**, Heiko Paulheim, Jan Portisch, Artem Revenko, Annette Ten Teije, and Frank Van Harmelen. 2023. Combining Machine Learning and Semantic Web: A Systematic Mapping Study. ACM Computing Surveys, 55(14s).

List of Figures

1.1 News consumption by source in Europe, 2013-2025 . . . 2
1.2 Proportion of respondents using each source of news, by market, in 2025 . . . 3
1.3 Number of Wikipedia pages over time . . . 8
1.4 Usage statistics of content languages for websites in August 2025 . . . 8
1.5 Digital language support per country . . . 8
1.6 Schematic overview of cross-lingual transfer paradigms . . . 9

2.1 Overview of the Transformer architecture . . . 21
2.2 Overview of BERT pretraining architecture . . . 26

3.1 Framework of the key components in personalized neural news recommenders . . . 32
3.2 An example of user modeling in personalized news recommendation . . . 37
3.3 Examples of user-news graphs used in graph-based user modeling . . . 41
3.4 Illustration of the NewsRecLib framework. . . . 53
3.5 A minimal configuration example for training an NRMS model in NewsRecLib. . . . 54

4.1 Illustration of the unified neural news recommendation framework . . . 70
4.2 Number of model parameters on MINDsmall . . . 73

5.1 Ranking performance of recommenders depending on the news encoder architecture and input features . . . 81
5.2 Retrieval and representational similarity of models with diferent news encoder architectures and input features . . . 82
5.3 Evolution of Jaccard similarity for different values of k . . . 83
5.4 Ranking performance of different recommenders depending on the user encoder architecture, for different base news encoder families . . . 83
5.5 Retrieval and representational similarity of models with different user encoder architectures . . . 84

6.1 Distribution of news over time and over outlets . . . 101
6.2 Corpora annotation and knowledge graph construction pipeline . . . 102
6.3 Distribution of outlets and news over political orientation classes . . . 102
6.4 Examples of words assigned to different sub-topics . . . 103
6.5 Example of an article's representation in NeMigKG . . . 105

7.1 Illustration of the MANNeR framewok . . . 116
7.2 Results of single-aspect customization for MANNeR and the best baseline on MIND-large . . . 121
7.3 t-SNE plots of the news embeddings in the test sets of MIND and Adressa . . . 121
7.4 Results of multi-aspect customization for MANNeR on MIND . . . 123
7.5 XLT in single-aspect customization, with MIND as target-language dataset . . . 124

8.1 ZS-XLT$_{\text{MONO}}$ ranking performance across the xMIND languages and English . . . 138
8.2 Relative percentage difference in ranking between ZS-XLT$_{\text{BILING}}$ and full English training and testing . . . 139
8.3 FS-XLT ranking performance with a *random replacement* strategy . . . 140
8.4 Relative percentage difference in ranking performance under FS-XLT$_{\text{MONO}}$ compared to ZS-XLT$_{\text{MONO}}$ with a *random replacement* strategy . . . 141
8.5 Relative percentage difference in ranking performance under FS-XLT$_{\text{BILING}}$ compared to ZS-XLT$_{\text{BILING}}$ with a *random replacement* strategy . . . 141
8.6 FS-XLT ranking performance wit a *category-based replacement* strategy . . . 142
8.7 Relative percentage difference in ranking performance under FS-XLT$_{\text{MONO}}$ compared to ZS-XLT$_{\text{MONO}}$ with a *category-based replacement* . . . 142
8.8 ZS-XLT$_{\text{MONO}}$ versus FS-XLT$_{\text{MONO}}$ ranking performance with *random* and *category-based* replacement strategies . . . 143
8.9 ZS-XLT$_{\text{MONO}}$ and FS-XLT$_{\text{MONO}}$ ranking performance for different fine-tuning strategies . . . 145
8.10 Cosine similarity between the embeddings of English and target-language news articles 147
8.11 Distribution of principal angles for the top $k = 50$ PCA subspaces, and Grassmann distance, for the embeddings of the English and the target-language news articles . 148
8.12 CKA similarity between the embeddings of English and target-language news articles 149
8.13 Annotator agreement for three types of translation aspects in the annotation task . . 150
8.14 Annotation task results for each type of translation aspect . . . 152
8.15 ZS-XLT$_{\text{MONO}}$ ranking performance in terms of MT system . . . 152
8.16 Ranking performance in terms of MT system . . . 153

9.1 Distribution of texts in PolyNews across languages and provenance datasets . . . 157
9.2 ZS-XLT performance in few-shot recommendation . . . 164
9.3 Ranking performance of LFRec-SCL with frozen NaSE embeddings versus LFRec-SCL with frozen LaBSE embeddings . . . 164

C.1 Example for logging the model size, number of trainable and non-trainable model parameters in NewsRecLib . . . 239

D.1 Number of model parameters on MINDlarge . . . 241

E.1 Effect of different (a) NE inputs and (b) model design/training choices on MANNeR's content-based personalization performance . . . 244
E.2 Results of single-aspect customization for MANNeR and the best baseline on Adressa 245
E.3 Results of multi-aspect customization for MANNeR on Adressa . . . 245
E.4 XLT in single-aspect customization, with Adressa as target-language dataset . . . 246

G.1 Distribution of texts across language pairs in PolyNewsParallel . . . 251
G.2 Recommendation performance of LF-Rec with frozen NaSE and LaBSE embeddings across the xMIND languages and English . . . 252

List of Tables

3.1 Overview of public news recommendation datasets . . . 50
3.2 Statistics of the Adressa and MIND datasets . . . 51
3.3 Comparison of NewsRecLib against existing open-source libraries that support neural news recommendation . . . 52
3.4 List of currently available models in NewsRecLib and supported configurations . . . 56

4.1 Recommendation performance of the compared models under combinations of click behavior fusion and training objectives . . . 72

5.1 Abbreviations and their descriptions . . . 78
5.2 Text encoder architectures . . . 78
5.3 Multi-feature aggregation strategies for combining textual and categorical representations of news . . . 79
5.4 User encoder architectures . . . 79

6.1 GeNeG knowledge graph statistics . . . 93
6.2 Statistics of the the German subset of x-stance . . . 93
6.3 Questions for stance detection . . . 94
6.4 Number of articles per stance and pairing question . . . 94
6.5 GeNeG user data statistics . . . 95
6.6 Click-through rate prediction results . . . 96
6.7 Average sentiment scores for users and recommenders . . . 97
6.8 Average stance scores for users and recommenders . . . 97
6.9 Recommender-user sentiment correlation counts . . . 98
6.10 Recommender-user sentiment bias correlation . . . 98
6.11 Recommender-user stance correlation counts . . . 99
6.12 Recommender-user stance bias correlation . . . 100
6.13 NeMig corpora statistics . . . 103
6.14 Statistics of extracted named entities . . . 104
6.15 Statistics of the German and English NeMigKG . . . 106
6.16 Excerpt of questionnaire used in the US online user study . . . 107
6.17 Recommendation performance of different NNRs . . . 109
6.18 Content personalization and aspect diversity performance of different NNRs . . . 110
6.19 Content and aspect personalization performance of different NNRs . . . 110
6.20 Content personalization and aspect-based diversity performance of DKN with different variants of NeMigKG . . . 111

7.1 Content-based recommendation performance . . . 119
7.2 Single-aspect diversification . . . 120
7.3 Single-aspect personalization . . . 122

8.1 The languages of xMIND . . . 133
8.2 Diversity indices of multilingual news recommendation datasets . . . 133
8.3 Statistics of a subset of Global Voices . . . 134
8.4 Hyperparameter optimization results on the subset of Global Voices . . . 134
8.5 Statistics of xMIND dataset splits . . . 134
8.6 ZS-XLT$_{\text{MONO}}$ recommendation performance . . . 137
8.7 ZS-XLT$_{\text{MONO}}$ recommendation performance with different fine-tuning strategies . 145

9.1 ZS-XLT recommendation performance with a fine-tuned news encoder . . . 162
9.2 ZS-XLT recommendation peformane with a frozen news encoder . . . 163

A.1 Overview of published resources . . . 231

B.1 Distribution of number of Wikipedia pages over time . . . 233

E.1 Number of model parameters . . . 243
E.2 Search spaces used for hyperparameter optimization and best values found for all models . . . 244
E.3 Inference time for the different NNRs on the test portions of the MINDlarge and Adressa datasets . . . 246

F.1 xMIND language support in mPLMs . . . 247
F.2 Hyperparameter optimization settings for xMIND experiments . . . 247

G.1 Languages of PolyNews . . . 250

List of Acronyms

ACC Accuracy. 95, 96

Adressa Adressa Dataset. 9, 49–52, 56, 60, 68, 91, 118–123, 130, 131, 172, 243–246

AI Artificial Intelligence. 2, 3, 5, 173, 174, 176

AUC Area Under the Curve. 46, 47, 57, 71–73, 79, 95, 96, 108, 109, 111, 118, 119, 136, 137, 144, 145, 161–163, 252

BCE Binary Cross-Entropy. 43, 46

BERT Bidirectional Encoder Representations from Transformers. 18, 25–27, 34, 93, 160

BPR Bayesian Personalized Ranking. 43, 46

CAUM News Recommendation with Candidate-aware User Modeling. 45, 46, 56, 69–73, 108–110, 118–120, 122, 135, 137, 140–143, 160–163, 243, 244, 246, 247, 251

CBOW Continuous Bag-of-Words. 19, 20, 27

CE Cross-Entropy. 43, 46, 69, 71–73, 160

CenNewsRec Privacy-preserving News Recommendation Model Learning. 45, 46, 56, 69, 71–73

CNN Convolutional Neural Network. 33, 34, 36, 38, 39, 42, 44, 45, 69, 78, 79, 81–84, 107, 108, 135, 160

CTR Click-Through Rate. 36, 43, 47, 57, 58, 96

DKN Deep Knowledge-aware Network for News Recommendation. 44, 46, 56, 69, 71, 72, 108–111

EF Early Fusion. 56, 68–73, 241

EL Entity Linking. 103, 104

F_1 F_1-score. 95, 96, 111

FS-XLT Few-shot Cross-Lingual Transfer. 9, 13, 14, 127–129, 134, 136, 137, 140–149, 151, 153

GloVe Global Vectors. 19, 20, 33, 45, 46, 109

GRU Gated Recurrent Unit. 38–40, 45, 69, 78, 79, 83–85, 108, 135, 160

IR Information Retrieval. 61, 172, 173

KG Knowledge Graph. 36, 37, 91–93, 101, 103–106

KGE Knowledge Graph Embedding. 36

LaBSE Language-agnostic BERT Sentence Embedding. 28, 61, 135–139, 159–164, 247, 249–252

LF Late Fusion. 11, 12, 56, 68–73, 78, 79, 83–85
LFRec Late Fusion-based Recommender. 160–164, 251, 252
LLM Large Language Model. 2, 4, 21, 27, 35, 41, 42, 52, 58, 59, 130, 131, 149, 172–174, 176
LM Language Model. 4, 8, 9, 13, 14, 17–19, 21–23, 25–28, 103, 155–158, 160–163
LSTM Long Short-Term Memory. 21, 28, 38, 39
LSTUR Neural News Recommendation with Long-and Short-term User Representations. 34, 45, 46, 56, 69–73, 117, 119, 120, 122, 135, 137–140, 142, 243, 244, 247

MANNeR Multi-Aspect Neural News Recommendation. 12, 56, 57, 113–124, 135, 137–148, 151–153, 160–163, 170, 171, 175, 176, 231, 243–247, 251
MIND MIcrosoft News Dataset. 9, 13, 14, 49–52, 56, 60, 68, 71–73, 80, 91, 109, 118–124, 129–132, 134–138, 141, 145, 146, 149–151, 153, 159, 160, 162, 163, 172, 173, 241, 243–246, 248, 252
MINER Multi-interest Matching Network for News Recommendation. 45, 46, 56, 57, 117, 119, 120, 122, 135, 137, 138, 141–143, 243, 244, 246, 247
MINS News Recommendation via Multi-Interest News Sequence Modelling. 45, 46, 56, 69–72, 108–110, 118–120, 122, 135, 137, 138, 141, 142, 160–163, 243, 244, 246, 247, 251
MLM Masked Language Modeling. 25–28, 62
MLP Multilayer Perceptron. 42
mPLM Multilingual Pretrained Language Model. 27, 46, 61, 62, 128–131, 135, 138, 139, 143–149, 247
MRR Mean Reciprocal Rank. 47, 57, 71–73, 79, 108, 109, 118, 119, 136, 137, 144, 145, 161–163
MSE Mean Squared Error. 43, 61
MT Machine Translation. 13, 62, 149, 151–153, 156, 159, 164

NAML Neural News Recommendation with Attentive Multi-view Learning. 34, 45, 46, 56, 69–72, 107–110, 117, 119–122, 135, 137–148, 151–153, 160–163, 243, 244, 246, 247, 251
NaSE News-adapted Sentence Encoder. 13, 80, 81, 155, 156, 158–165, 171, 176, 231, 250–252
NCE Noise Contrastive Estimation. 44, 46
NDCG Normalized Discounted Cumulative Gain. 47, 57, 71–73, 79, 80, 108–111, 118–120, 122, 136–145, 151–153, 161–164
NE News Encoder. 67–69, 72, 73, 116, 124, 162, 243, 244, 251
NER Named Entity Recognition. 103, 104
NewsRecLib Neural News Recommendation Library. 52–57, 81, 129, 136, 161, 231, 235–239, 244, 251
NLL Negative Log-Likelihood. 43, 46
NLP Natural Language Processing. 4, 17–19, 21, 25, 28, 33, 61, 172, 173
NMT Neural Machine Translation. 128, 129, 132
NN Neural Network. 4
NNR Neural News Recommender. 5, 7, 10–14, 31, 32, 34, 35, 38, 40, 42–47, 51–58, 60–62, 67–71, 74–81, 84–86, 90, 107, 109, 110, 113–119, 124, 127, 129, 130, 134–141, 144, 146–149, 151, 153, 155, 156, 159–163, 169–173, 231, 243, 246–248, 251
NPA Neural News Recommendation with Personalized Attention. 34, 44, 46, 56, 69, 71, 72

NRMS Neural News Recommendation with Multi-head Self-attention. 34, 39, 44–46, 54, 56, 69, 71, 72, 107–110, 117, 119, 120, 122, 243, 244, 246
NRS News Recommender System. 3, 6, 10, 13
NSP Next Sentence Prediction. 25–27

OOV Out-of-Vocabulary. 18–20, 22

PLM Pretrained Language Model. 12, 25, 27, 34, 35, 46, 52, 54, 56, 57, 78, 80–82, 85, 114–117, 123, 124, 155, 243, 246
PolyNews PolyNews Dataset. 13, 155–159, 161, 164, 165, 231, 249, 250
PolyNewsParallel PolyNewsParallel Dataset. 13, 155, 156, 158, 159, 161, 164, 165, 231, 251

RNN Recurrent Neural Network. 4, 39, 40, 45
RoBERTa Robustly Optimized BERT Approach. 18, 25–28, 34, 54, 80, 118, 243
RS Recommender Systems. 1

SCL Supervised Contrastive Loss. 69, 71–73, 160
SE Sentence Encoder. 27–29, 78, 80, 82–84, 153, 155–163
SentiDebias Sentiment-Debiased Recommender System. 46, 56, 108–110, 118–120, 122, 243, 244, 246
SentiRec Sentiment Diversity-aware News Recommender. 44, 46, 56, 118–120, 122, 243, 244, 246
SG Skip-Gram. 19, 20, 27

TANR Neural News Recommendation with Topic-aware News Representation. 44–46, 56, 108–110, 118–122, 135, 137–140, 142, 143, 243, 244, 246, 247
TF-IDF Term Frequency–Inverse Document Frequency. 33, 94, 96–100
TLM Translation Language Modeling. 28
TransD Knowledge Graph Embedding via Dynamic Mapping Matrix. 36, 44, 109, 110
TransE Translating Embeddings for Modeling Multi-Relational Data. 36, 49, 71, 135, 243
TSDAE Transformer-based Sequential Denoising Auto-Encoding. 158, 159

UE User Encoder. 67–69, 71–73, 244

XLM-RoBERTa XLM-RoBERTa. 9, 27, 28, 56, 61, 102, 103, 118, 133, 135–141, 144–149, 160–163, 247, 251
XLT Cross-Lingual Transfer. 9, 123, 124, 156, 157, 160, 163, 245, 246
xMIND Multilingual MIcrosoft News Dataset. 12, 13, 52, 56, 127, 129, 131–140, 142, 143, 145–151, 153, 159–164, 171, 173, 176, 231, 247, 252

ZS-XLT Zero-shot Cross-Lingual Transfer. 9, 12–14, 127, 129, 134, 136–145, 151–153, 155, 156, 161–165

Referring to specific languages, we use ISO 639-3 codes, e.g., ENG for English.

CHAPTER 1

INTRODUCTION

There are things known and there are things unknown, and in between are the doors of perception.

ALDOUS HUXLEY

In this Chapter, we first motivate the need for personalized neural recommender systems in the news domain, advocating for models that can support diverse recommendation scenarios across news aspects and natural languages (Section 1.1). Afterwards, we summarize our main contributions (Section 1.2) and conclude with an outline of this thesis (Section 1.3).

1.1 Motivation

Recommender systems (RS) are broadly defined as tools and techniques that filter vast amounts of information and *suggest items* deemed potentially *relevant* (i.e., interesting and useful) to a particular user (Resnick and Varian, 1997; Burke, 2002; Ricci et al., 2022). Their primary goal is to combat information overload from an overwhelming array of choices, such as selecting one restaurant among hundreds in a city, and to enhance user engagement and satisfaction (Ricci et al., 2022). In most cases, recommendations are *personalized*, meaning that they are tailored to the individual interests of users (Jannach et al., 2010). Personalized recommender systems estimate the relevance of an item based on a user's preferences and past behavior. Recommendation algorithms infer these preferences from large datasets of user interactions, known as *interaction logs*. These logs contain either explicit feedback (e.g., user-provided ratings) or implicit feedback (e.g., browsing history). The most common approaches include collaborative filtering (i.e., recommending items favored by users with similar preferences), content-based filtering (i.e., suggesting items similar to those that the user has liked before), and hybrid methods which combine both strategies (Adomavicius and Tuzhilin, 2005).

Recommender systems are nowadays ubiquitous, deeply integrated into the digital platforms and applications that shape modern life. They guide our choices in listening to music, discovering new movies, shopping for goods, ordering food, reading news, connecting on social media, or making travel plans. The scale of their impact is extensive. An estimated 80% of viewing hours on Netflix for its over 3 billion subscribers are driven by its recommendation engine (Gomez-Uribe and Hunt, 2015).[1] Similarly, approximately 70% of YouTube's video consumption originates from algorithmic suggestions (Solsman, 2018), while roughly 35% of purchases on Amazon are directly attributable to its product recommendations (MacKenzie et al., 2013). As a result, recommender systems now profoundly influence decision-making across nearly every digital domain, from entertainment (Covington et al., 2016; Goodrow, 2025; De Nadai et al., 2024, *inter alia*) and services (Chaudhari and Thakkar, 2020; Bunnell et al., 2020; Etemadi et al., 2023, *inter alia*), to social networking (Kenthapadi et al., 2017; Geyik et al., 2018; Twitter, 2023) and e-commerce (Smith and Linden, 2017;

[1] https://ir.netflix.net/ir-overview/profile

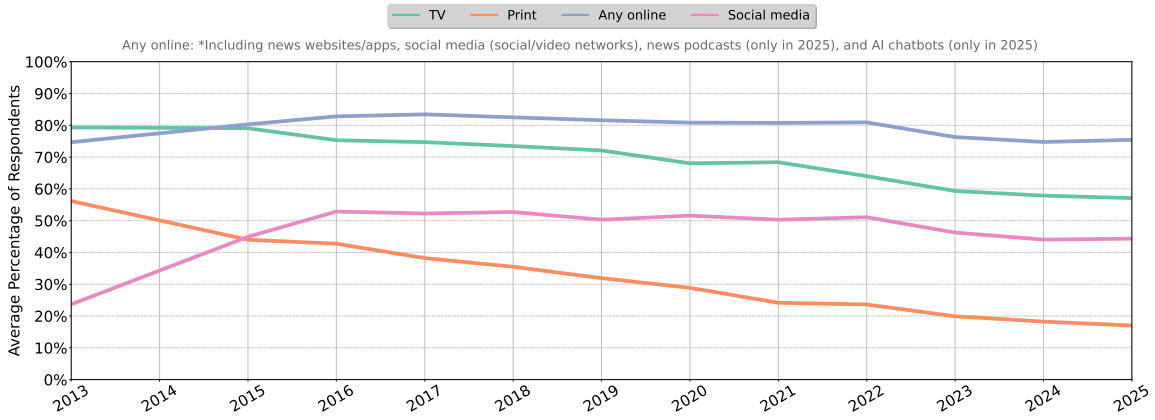


Figure 1.1: News consumption by news source in Europe, between 2013 and 2025. We show the percentage of respondents, averaged over the 24 countries included in the available data from the Reuters Digital News Report 2025 (Newman et al., 2025).[2]

Chen et al., 2019; Wang et al., 2024c, *inter alia*). Notably, in many social media platforms, such as Instagram, TikTok, or Twitter / X, user feeds are now driven almost entirely by recommendation algorithms (TikTok, 2020; Narayanan, 2023; Vorotilov and Shugaepov, 2023).

Recommender Systems for News. Arguably, recommender systems shape individuals' worldview through the way in which they filter information and expose only a small subset of available content. This influence becomes particularly critical in the domain of news, where news media serve as a cornerstone of democratic society by keeping the public informed with timely and relevant information, and by fostering public discourse (Balkin, 2017; Helberger, 2021). Online news consumption is among the top uses of the Web, after communication and e-commerce (Eurostat, 2024). For example, data from the European Union shows that in 2024, 65% of individuals accessed news online, far surpassing the 49% who used commercial video-on-demand services (Eurostat, 2024).

This reflects a notable shift in how audiences access and engage with news. Over the past decade, the media landscape has seen a sharp decline in the usage of traditional news outlets, particularly print newspapers and TV, and a concurrent rise of online media, in the form of news websites, aggregation platforms, and social media. As illustrated in Fig. 1.1, in Europe, print newspaper readership has plummeted from 56% in 2013 to only 17% in 2025, while online news consumption has consistently remained above 75% (Newman et al., 2025). This trend is mirrored across continents, where online sources remain dominant and print is marginal, as indicated by Fig. 1.2. More recently, the advent of generative large language models (LLMs), such as ChatGPT (OpenAI, 2022; OpenAI et al., 2024) or Llama 3 (Touvron et al., 2023), has introduced another paradigm shift in the media landscape, with AI-driven chat interfaces emerging as a new source for news, used by approximately 7% of users (15% for those under 25) in 2025. In parallel, search engines like Google are increasingly evolving into hybrid AI-search platforms, combining traditional link-based retrieval with generative summaries and recommendations, a shift that is likely to further reshape how audiences find and engage with news (Reid, 2023; Kiskola et al., 2025).

[2] Fig. 1.1 is based on data provided by the Reuters Institute (Newman et al., 2025). The 24 European countries considered, together with the year for which data was first available, are: United Kingdom (2013), Austria (2015), Belgium (2016), Bulgaria (2018), Croatia (2017), Czech Republic (2015), Denmark (2013), Finland (2015), France (2013), Germany (2013), Greece (2016), Hungary (2016), Italy (2013), Ireland (2015), Netherlands (2015), Norway (2016), Poland (2015), Portugal (2015), Romania (2017), Slovakia (2017), Spain (2013), Sweden (2016), Switzerland (2016), and Turkey (2015).

[3] Fig. 1.2 covers the following countries in each of the four markets included in the Reuters Digital News Report 2025 (Newman et al., 2025): Africa – Kenya, Morocco, Nigeria, South Africa; Americas – United States of America, Argentina, Brazil, Canada, Chile, Colombia, Mexico, Peru; Asia Pacific – Australia, Hong Kong, India, Indonesia, Japan, Malaysia, Philippines, Singapore, South Korea, Taiwan, Thailand; Europe – United Kingdom, Austria, Belgium, Bulgaria, Croatia,

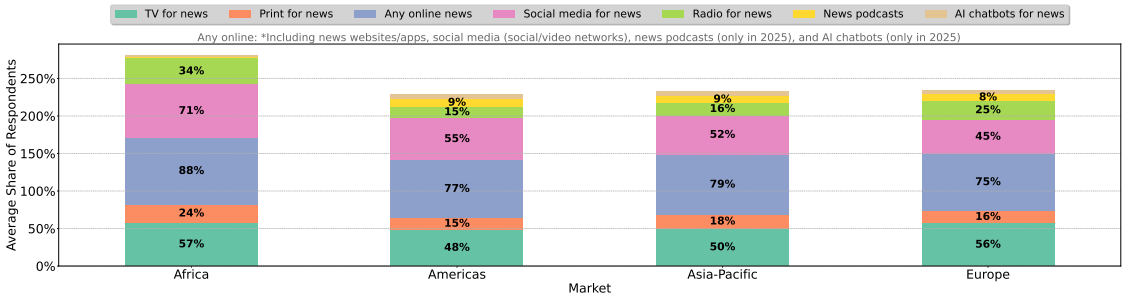


Figure 1.2: Proportion of respondents using each source of news in a week, by market, in 2025. The percentage of respondents is averaged over the countries included in the respective market according to the data provided by the Reuters Digital News Report 2025 (Newman et al., 2025).[3]

1.1.1 Personalized Neural News Recommendation

In today's digital news landscape, human editors are no longer the sole gatekeepers of information. Instead, *news recommender systems* (NRS) have become indispensable for filtering the sheer volume of articles published daily, personalizing content, and curating the information users see on news websites, aggregators, or social media platforms (Das et al., 2007; Kunert and Thurman, 2019; Møller, 2022; Piscopo et al., 2024, *inter alia*). Reflecting this central role, a recent Reuters Institute survey of 326 media leaders revealed that 80% consider artificial intelligence (AI) important for news distribution and recommendation, including personalized homepages and alerts (Newman and Cherubini, 2025). The ubiquity of personalized algorithmic curation is evident also among audiences, with nearly half (49%) expressing comfort with the automatic selection of content. This acceptance is primarily motivated by the desire to access more relevant news while saving time and effort, although it remains slightly lower than in domains such as music or movies, where acceptance levels exceed 50% (Newman et al., 2025). These trends coincide with a broader shift in how audiences access news: direct visits to news websites have declined from 33% in 2018 to 22% in 2024, while search engines and aggregators (33%) now serve as a more common gateway than social media (29%) or visiting news outlets directly (22%) (Newman et al., 2024). This increasing reliance on aggregated and algorithmically curated feeds, rather than deliberately visiting a small set of news websites, further highlights the relevance of personalized recommender systems in the news domain.

We define an *online news platform* as any digital medium that delivers news content to users, either via an outlet's website or application (e.g., *spiegel.de*, *nytimes.com*) or via aggregator or social media platforms (e.g., Google News, Apple News, TikTok). Unlike publisher websites, which often present the same landing page to all users, *algorithmically curated platforms* personalize the feed by adapting the selection and ordering of content to individual user preferences. Our focus specifically lies on such platforms that implement news recommender systems, dynamically selecting and ranking content for each user. Throughout this thesis, we use the term *news* to refer exclusively to *textual news articles*, namely written pieces comprising titles and bodies, as opposed to multimodal formats such as news videos, podcasts, or audio summaries.

When a user visits an online news platform, a personalized news recommender system will retrieve a small set of candidate articles from a large news pool, ranking them according to the inferred user interests, and presenting the top-K results (Wu et al., 2023). User interactions, such as clicks or reading time, are then recorded to update their profiles. While this resembles workflows in other domains, the unique characteristics of news demand a different approach, one that relies on *personalized neural recommendation models*.

Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland, Turkey.

The Need for Personalization. On the one hand, the news domain is defined by its time sensitivity and volume. The 24/7 news cycle produces a constant stream of news articles whose relevance decays rapidly (Li et al., 2011a; Özgöbek et al., 2014; Karimi et al., 2018). This high churn rate makes manual curation nearly impossible and places a heavy burden on readers, who require additional time and effort to stay engaged and find relevant information. While early, non-personalized systems based on factors like location (Asikin and Wörndl, 2014; Garrido et al., 2015; Chen et al., 2017), time (Park et al., 2017b; Lommatzsch et al., 2017), or popularity (Doychev et al., 2015; Corsini and Larson, 2016; Ludmann, 2017) offered a basic filter, they nevertheless exhibited various biases. Such systems can inadvertently amplify clickbait or misinformation, for instance when popularity- or time-based recommenders overemphasize articles that attract short-term bursts of attention due to controversy (Klimashevskaia et al., 2024; Yang, 2016). They may also foster localized echo chambers: popularity-based algorithms tend to suppress niche content or smaller providers, location-based models may reinforce region-specific narratives through geographically clustered exposure, and demographic-based recommenders can amplify the perspectives of particular communities (Bastos et al., 2018; Sánchez et al., 2023; Klimashevskaia et al., 2024). Moreover, non-personalized approaches generally underperform in driving user engagement, as they do not align content recommendations with individual user preferences (Raza and Ding, 2022). Effective news recommendation, therefore, hinges on *personalization* to understand and cater to the diverse and specific interests of each user.

The Need for Neural Approaches. On the other hand, news articles contain rich, unstructured, textual information in their titles and bodies. Understanding and modeling their semantic meaning and context is crucial for accurate recommendations. This inherently requires advanced natural language processing (NLP) techniques, making neural networks (NNs) the de facto tool for learning deep representations of news content (Raza and Ding, 2022; Wu et al., 2023). The evolution of news recommender systems has been, in fact, closely intertwined with advances in representation learning methods (detailed in Chapter 2): from static and contextualized word embeddings, to pretrained language models (LMs), and most recently, large language models (LLMs).

Similarly, the user-side dynamics in the news domain demand advanced neural architectures. Users rarely provide explicit feedback on news websites and aggregators, and, unlike, e.g., in online retail, here there is no difference between *viewing* and *purchasing an item* (Iana et al., 2024). Instead, interest is inferred from implicit signals like clicks and reading time (Liu et al., 2010b; Wu et al., 2023). Moreover, users' interests are diverse and evolve dynamically over time (Harandi and Gulla, 2015; Gulla et al., 2014; Feng et al., 2020). A person may have long-term interests (e.g., climate change) motivated by their socio-economic and personal background, alongside short-term preferences, highly affected by breaking news and context (e.g., a local election, a tech launch) (Harandi and Gulla, 2015; An et al., 2019). This behavior, combined with the high item churn of news (Das et al., 2007) and often session-based user interactions (Doychev et al., 2015; Feng et al., 2020), exacerbates the classic cold-start problem (i.e., no interaction data for new items or users) and data sparsity (i.e., interaction data only for a small subset within a large pool of articles) (Karimi et al., 2018). Architectures such as recurrent neural networks (RNNs) and Transformers (see Section 2.1) are thus essential for modeling the dynamic and sequential patterns of a user's reading history, as they can accurately capture short-term intent and contextual shifts (Li and Wang, 2019; Wu et al., 2023).

Overall, this goes to show that the time-sensitive, high-churn nature of news, coupled with the richness of its content and the dynamics of user interests, make personalized neural approaches essential for accurate news recommendations. Yet, these technical demands are only one side of the story. An effective and responsible news recommender system must also grapple with the societal and linguistic challenges of algorithmic news curation.

1.1.2 The Case for Multifaceted Neural News Recommendation

Individuals differ widely in their demographics, socio-economic background, spoken languages, and personal concerns. These differences naturally translate into diverse needs and preferences when seeking information. Personalized news recommender systems profoundly impact how users interact with news content, as well as the content they encounter. While such systems can effectively drive engagement, they risk narrowing users' exposure to information when optimized solely for clicks and relevance (Freedman and Sears, 1965; Pariser, 2011; Nguyen et al., 2014; Bakshy et al., 2015). Moreover, they may inadvertently amplify biases (Bozdag, 2013; Mansoury et al., 2020; Klimashevskaia et al., 2024) and misinformation (Fernandez et al., 2024; Jakobsen et al., 2025), stemming from both societal imbalances in the training data and algorithmic design choices (Bandy and Diakopoulos, 2020; Mitchell et al., 2020; Liu et al., 2021b; Einarsson et al., 2025, *inter alia*).

A responsible news recommender system should, therefore, balance predictive accuracy and engagement maximization objectives with broader human-centered and societal values (Kazienko and Cambria, 2024; Said, 2024; Seddik, 2024). It should foster exposure to diverse perspectives, limit the amplification of bias, misinformation, and sensationalism, respect user autonomy, ensure access to trustworthy information, and maintain transparency and accountability in how recommendations are generated (Elahi et al., 2022; Milano et al., 2020; Bernstein et al., 2021; Deldjoo et al., 2024b). Responsibility, in this sense, extends beyond technical performance to encompass the ethical and societal implications of algorithmic decision making, requiring systems to be transparent, fair, and accountable for their influence on individual users and public discourse (Vrijenhoek et al., 2021; Tintarev et al., 2024; Eskens, 2022; Zhao et al., 2025a). This aligns with the European Commission's *Framework for Ethical AI*, which defines trustworthy AI as ensuring human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, and accountability (European Commission, 2019).

A central dimension of inclusivity in personalized news recommendation is language. Despite growing online linguistic diversity, most AI technologies, including recommender systems, are developed for a handful of high-resource languages such as English or Chinese. This inequitable distribution of technologies across languages gives rise to the *digital language divide*, a considerable gap between those who have, and those who do not have, access to digital technology (National Telecommunications and Information Administration, 2001; Van Dijk, 2020). Such inequities restrict equitable access to information and exacerbate global disparities in digital participation.

Together, the challenges of biased algorithmic news curation and the digital language divide affect not only the breadth and diversity of information to which individuals are exposed, but also their very access to that information. The right to "seek, receive and impart information and ideas through any media and regardless of frontiers", as enshrined in the United Nations' Universal Declaration of Human Rights Article 19 (United Nations, 1948), is directly undermined by linguistic barriers. Echoing this, the European Union's Declaration on Digital Rights emphasizes that "everyone should have access to a trustworthy, diverse and multilingual digital environment", as an essential condition for "pluralistic public debate and effective participation in democracy in a non-discriminatory manner." (European Parliament, Council of the European Union, European Commission, 2023). These principles underscore that news recommendation technologies must "leave nobody behind" (European Parliament, Council of the European Union, European Commission, 2023), ensuring equitable access to trustworthy and diverse information for all users, regardless of language, demographic background or socio-economic status.

In this thesis, we argue that building truly inclusive and responsible personalized neural news recommender (NNR) systems requires a *multifaceted approach*: one that moves beyond relevance-only optimization to integrate two core dimensions: accounting for the *societal implications of algorithmic news curation*, and mitigating the *inequities created by linguistic divide*. The following sections examine these two challenges in turn: first, the effects of algorithmic news curation on users' information diets, and second, the role of language in shaping digital access. These discussions lay

the groundwork for the challenges addressed and contributions of this thesis.

Effects of Algorithmic News Curation. Online news readers have diverse information needs and worldviews, yet they tend to prefer information in agreement with their prior opinions (Freedman and Sears, 1965). To maximize user satisfaction, personalized recommenders are typically optimized for similarity, either to the user's past behavior or to that of similar users (Jannach et al., 2010).

Information Diversity. Relevance maximization often reinforces exposure to content already consumed, filtering out articles deemed irrelevant or counter-attitudinal (Ludwig et al., 2023). Such mechanisms raise concerns about algorithmically induced "filter bubbles" (Pariser, 2011) and "echo chambers" (Sunstein, 2001), namely increasingly personalized content ecosystems in which users encounter mainly attitude-congruent or ideologically consistent perspectives and limited exposure to alternatives. This inherent control of information exposure by recommendation algorithms is especially critical in the news domain, due to the pivotal role of news media in democratic societies (Baker, 2006; Helberger, 2021). In contexts such as politics, conflict, public health, or climate change, exposure to only one side of a debate can foster attitude-reinforcing spirals (Donsbach and Mothes, 2013; Stroud, 2017), reduce interaction across ideological groups (Bright, 2018), and influence opinion formation and voting behavior (Bartels, 1993; Dewenter et al., 2019). Over time, these dynamics can fuel polarization, degrade civil discourse, and even contribute to political or ideological radicalization (Jensen et al., 2012; Boutyline and Willer, 2017; Barberá, 2020; Liu et al., 2021b; Whittaker et al., 2021; Ludwig and Müller, 2022).

Empirical evidence on the prevalence of filter bubbles in news recommender systems is, however, mixed (Ludwig and Müller, 2022). For instance, Bakshy et al. (2015) found that algorithmic ranking reduced exposure to cross-cutting content in terms of ideologies by 15% among 10.1 million US Facebook users, while González-Bailón et al. (2023) showed increased ideological segregation between liberal and conservative audiences during the 2020 US election due to algorithmic news curation. A large-scale audit of Twitter / X similarly reports asymmetric amplification, particularly of right-leaning content (Huszár et al., 2022). Liu et al. (2021b) further demonstrate that different news recommendation algorithms give rise to distinct forms of filter bubbles, thereby contributing to a more homogeneous set of viewpoints even among individuals with otherwise divergent preferences. Their findings also indicate that users with more extreme predispositions receive recommendations that are less diverse in both topical coverage and political perspectives, precisely because these recommendations tend to yield higher click-through rates (Liu et al., 2021b).

Conversely, several studies find limited or no evidence of filter bubbles in news recommenders (Fletcher and Nielsen, 2017). Möller et al. (2018), for example, discovered that algorithmic recommendations for a Dutch newspaper were as diverse as journalistic ones. Similarly, Haim et al. (2018) found that personalization on Google News had no effect on source diversity, whereas explicit personalization had only a minor impact on content diversity, modestly increasing the number of articles aligned with users' existing topical preferences. Flaxman et al. (2016) reported that algorithmic curation simultaneously widened ideological gaps while increasing exposure to content from other sides of the political spectrum. However, many of these studies stem from survey data or simulated settings rather than real-world systems (Ludwig et al., 2023). These inconclusive findings underscore both the methodological challenges of studying real-world recommender systems and the active role of news platforms in shaping content diversity. Some providers have introduced interventions to counteract narrowing effects of NRS, such as promoting diverse sources (Lurie and Mustafaraj, 2019; Piscopo et al., 2024) or ensuring demographic content diversity (BBC, 2022). Taken together, this body of work highlights the need to better understand not only whether filter bubbles emerge, but also how they affect content diversity and user outcomes, such as ideological polarization.

Algorithmic Bias and Fairness. Beyond information diversity, a second line of concern relates to algorithmic bias and fairness. Since recommenders learn from historical data, they risk perpetuating existing societal biases tied to race, gender, or political orientation (Stinson, 2022; Chen et al., 2023a;

Ekstrand et al., 2022; Wang et al., 2023). They may also disproportionately favor certain types of content or users, leading to unequal representations of demographic groups, topics, or article characteristics (e.g., sentiments, political orientation) (Chen et al., 2023a; Haim et al., 2018; Alam et al., 2022). For example, news recommenders often prioritize "soft news" (e.g., entertainment, lifestyle, sports) over "hard news" (e.g., politics, education, economy), and may underrepresent certain issues, topics, or political actors (Bandy and Diakopoulos, 2020; Einarsson et al., 2025). This can have serious consequences if audiences are repeatedly exposed to a narrow set of voices or viewpoints, particularly in terms of political participation (Bartels, 1993; Dewenter et al., 2019). Studies also report systemic biases toward certain outlets, often conservative ones, in systems like Google News (Haim et al., 2018). Such imbalances threaten diversity of viewpoints and sources, which are vital for social cohesion and pluralistic democracies (Bernstein et al., 2021).

To address the effects of algorithmic news curation, prior work has explored enhancing the diversity of recommendations along several dimensions, including topic (Raza and Ding, 2021; Shi et al., 2022; Wu et al., 2022c; Choi et al., 2022), sentiment (Wu et al., 2020c, 2022h), and viewpoint (Shivaram et al., 2022), as well as mitigating gender (Wu et al., 2021d) and popularity (Gharahighehi et al., 2021) biases. While these approaches help alleviate specific issues, they are often rigid when applied across multiple dimensions simultaneously (cf. Chapter 7). This is because they typically require retraining the underlying recommender whenever the recommendation objective changes (e.g., increasing topical diversity while reducing personalization). Such inflexibility limits their practicality in real-world scenarios, where user needs vary substantially; for instance, one user may prioritize exposure to diverse perspectives, while another may be interested less in sentiment diversity.

Although user modeling components in NNRs learn latent preferences from historical behavior, they are generally trained to optimize a single, global objective (e.g., relevance, diversity), and do not inherently distinguish between multiple dimensions of diversity or fairness at individual level. As a result, a model optimized for one objective, such as topical diversity, may not satisfy a user whose priority is viewpoint diversity, and vice versa. In other words, standard NNRs cannot dynamically adapt to varying user-specific trade-offs between multiple dimensions without retraining or fine-tuning. Addressing this limitation requires models capable of flexible, user-controllable recommendations, where trade-offs across aspects such as relevance, topical coverage, sentiment polarity, and viewpoint diversity can be adjusted per user simultaneously.

In sum, algorithmic news curation not only shapes what individuals read, but also who is heard. Left unchecked, it risks narrowing and homogenizing news diets, marginalizing culturally relevant perspectives, and reinforce structural inequalities. These risks emphasize why responsible and inclusive news recommendation technologies must target not only relevance and engagement, but also the societal impact of algorithmic curation.

Digital Language Divide. The global expansion of the Internet has greatly increased linguistic diversity online (Zuckerman, 2008). Language is therefore not just a medium of communication, but a key axis of inequality in access to personalized technologies. Wikipedia, one of the most visited websites, illustrates this trend: from a near-monolingual platform in 2001, it now hosts approximately 210 million pages in over 350 languages (see Appendix B). During this time (cf. Fig. 1.3), the relative share of English-written content has shrunk from 89.2% (i.e., about 20K of the 22K pages) in 2021 to 23.5% in 2025 (49M out of 210M pages). Yet, English remains the web's lingua franca. As shown in Fig. 1.4, the majority of online content remains concentrated among a few high-resource languages, with English being used in nearly half of the websites whose content language is tracked by W3Techs (2025), while most of the other languages appear in less than 0.1% of the web pages. The inequality is stark when considering the world's over 7,000 living languages and the number of native speakers (Eberhard et al., 2025): approximately half have no digital footprint, and a mere 33 (under 0.5%) of the remaining languages that are digitally supported to varying degrees (cf. Fig. 1.5) attain full digital capabilities (Kornai, 2013; Simons et al., 2022).

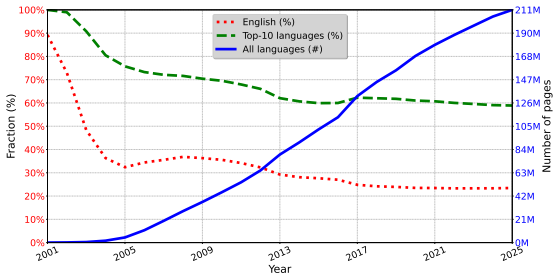


Figure 1.3: Number of Wikipedia pages, fraction of English pages, and fraction of pages of the ten largest Wikipedias each year, between 2001 and 2025.[4]

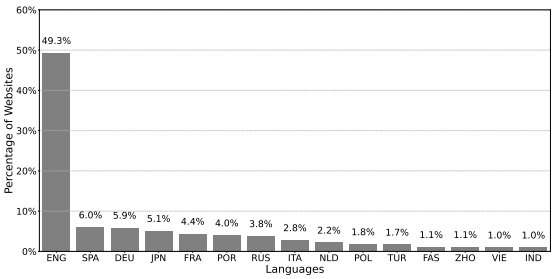


Figure 1.4: Usage statistics of content languages for websites in August 2025, according to W3Techs (2025). We report the percentage of websites using various content languages.

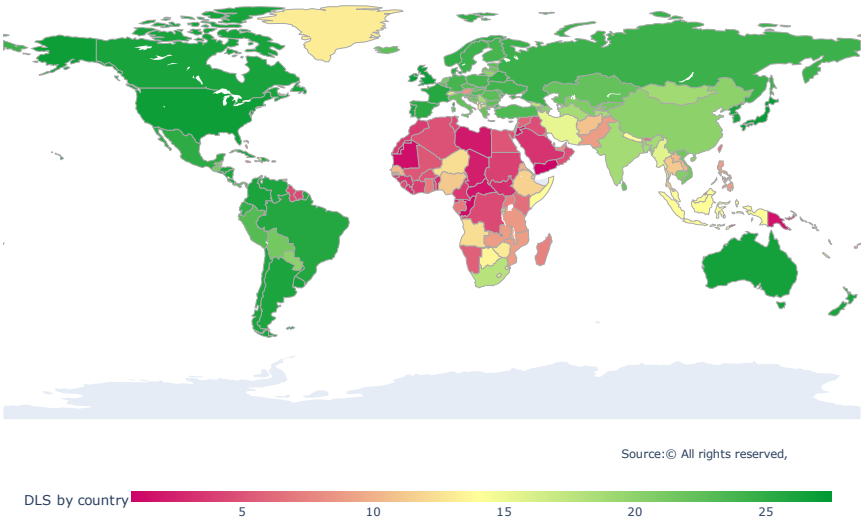


Figure 1.5: Digital Language Support (DLS) per country. Green indicates more DLS, red indicates less DLS (Derivation, 2022). DLS captures how well a given language is supported by digital tools and technologies (e.g., news sites, keyboards, spell checking, operating system, machine translation, speech tools, virutal assistants). DLS scores range from 0 to 28 for a given language, and are computed based on the type and number of tools supporting each language (Simons et al., 2022).[5]

Structural Divide. This imbalance translates directly into available data and technology. News recommenders rely on LMs, pretrained on web-scale corpora with trillions of tokens, to represent the content of articles. However, English makes up 80% of their training data (Touvron et al., 2023; Team et al., 2024; Bai et al., 2025). A recent large-scale study of nearly 400 datasets spanning 443 tasks, 608 languages, and 67 countries, shows that linguistic and geographic representation of data has not improved over the past decade, remaining predominantly western-centric, despite the significant increase in its total representation (Longpre et al., 2025). Even widely used or the latest multilingual

[4] Fig. 1.3 is adapted from Wikimedia Commons (2020).

[5] Simons et al. (2022) evaluated 143 digital tools across seven categories of digital language technology, ranging from the most to the least commonly supported: content (e.g., news sites), encoding (e.g, keyboards), surface (e.g., spell checking), localized (e.g., browser), meaning (e.g., machine translation), speech (e.g., speech-to-text), assistant (e.g. virtual assistant). They recorded the languages supported by each tool, producing a binary matrix of tool-language coverage. Within each category, a subscale score was computed based on how many tools support the language. The total score is obtained by summing the subscale scores across the seven categories. For more details on the computation of DLS scores, we refer the reader to Simons et al. (2022).

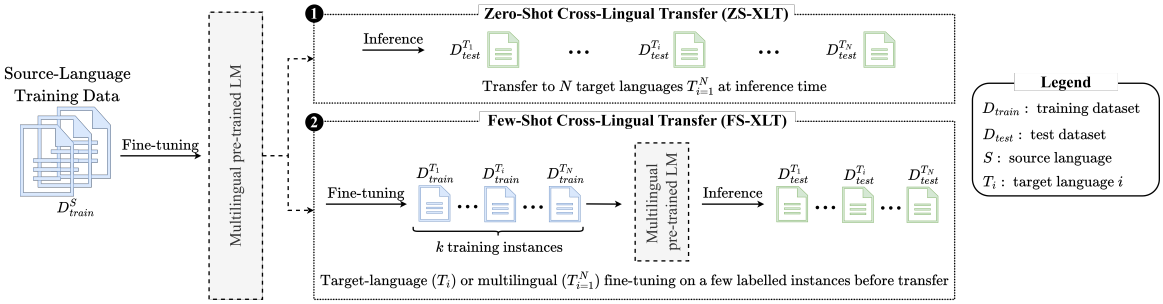


Figure 1.6: Schematic overview of cross-lingual transfer (XLT) paradigms, differentiating between zero-shot cross lingual-transfer (ZS-XLT) and few-shot cross-lingual transfer (FS-XLT).

pretrained LMs, such as XLM-RoBERTa (Conneau et al., 2020), mT5 (Xue et al., 2021), Gemma 3 (Team et al., 2025) or Qwen 3 (Yang et al., 2025), support only around 100-140 languages. Coverage is also skewed: high-resource languages not only benefit from higher-quality training data, but also from 10^5 more tokens than low-resource languages such as Georgian or Xhosa (Xue et al., 2021).[6] Consequently, LMs excel in high-resource languages, while downstream performance deteriorates proportionately to data scarcity and typological distance from English for low-resource languages (Lauscher et al., 2020; Ebrahimi et al., 2022; Wendler et al., 2024).

This disparity forms the *digital language divide*: speakers of high-resource languages benefit from advanced information technologies, while those with low-resource languages (e.g., Zulu, Somali, Quechua, Urdu, Irish) face degraded performance or no access at all (Eberhard et al., 2025). Unsurprisingly, existing news recommender systems and datasets are centered around a handful of resource-rich languages. Benchmarks are monolingual, e.g., MIND (Wu et al., 2020e) in English, Adressa (Gulla et al., 2017) in Norwegian, or EB-NeRD (Kruse et al., 2024) in Danish, or, when multilingual, still limited to well-represented languages (Wu et al., 2021b, Chapter 6).

Yet demand for multilingual access is rising. According to the Reuters Institute, 24% of online news readers want articles translated into their native language (Newman et al., 2025). Bridging this linguistic gap, however, remains a major challenge for personalized news recommendation, as collecting sufficient news or user-article interaction data for each language is rarely feasible. One promising strategy to mitigate this data scarcity is *cross-lingual transfer*, which enables models trained in high-resource languages to generalize to low-resources ones.

Cross-lingual transfer (XLT) is a form of transfer learning that leverages shared representations across languages to transfer knowledge from a source (typically high-resource) language to a target (generally low-resource) language, thereby improving performance in the latter (Devlin et al., 2019; Liu et al., 2019b; Conneau et al., 2020, *inter alia*). In essence, a multilingual LM – which may serve as the backbone of a recommender system – is first fine-tuned on large-scale source-language corpora, most commonly dominated by English. Two main paradigms of cross-lingual transfer are usually distinguished: (i) *zero-shot cross-lingual transfer* (ZS-XLT), where inference is performed directly with the multilingual LM without additional fine-tuning on target-language instances, and (ii) *few-shot cross-lingual transfer* (FS-XLT), where the LM is firstly fine-tuned on small amounts of target-language data before inference. Fig. 1.6 provides a visual overview of these XLT paradigms.

In the context of news recommendation, this means that a system trained on an English dataset such as MIND could, in principle, recommend articles in languages like Georgian or Swahili. However, in ZS-XLT model performance degrades sharply for languages that are typologically distant from English or absent from the pretraining corpora (cf. Chapters 8 and 9), while FS-XLT

[6]There are multiple definitions 'low-resource', 'resource-poor', or 'underrepresented' languages (Joshi et al., 2020; Yong et al., 2023; Nigatu et al., 2024). Here, we use the term to denote languages with a comparatively small digital footprint and limited representation in the web-scale training data of LMs.

requires labeled target-language data, which remains scarce and costly to obtain, particularly for underrepresented languages. Moreover, most NNRs rely on task-specific user interaction data, which are rarely available in low-resource settings or in cold-start scenarios (Wu et al., 2024c).

Consequently, the lack of language technology leaves many communities either entirely excluded from personalized news delivery or with only inequitable access. Users are often forced to consume news in a foreign language or to rely on machine translation, if such tools are even available, and thus, become marginalized in the global digital conversation.

Behavioral Divide. Language not only determines the content that individuals *can access*, but also shapes the information they *wish to access*. Today, many online readers are multilingual, speaking and consuming news in two or more languages (Ling et al., 2020, Chapter 8). For instance, over 74% of working-age adults in the European Union know at least one foreign language[7], 22% of US residents speak a non-English language at home (Patricia Ramos, 2025), and South Africans use 2.84 languages on average.[8] Among US Hispanics, about a quarter consume news equally in both English and Spanish (Naseer et al., 2024b). Globally, over half of the population speaks at least two languages (Grosjean, 2010). Multilingual users do not simply access information in different languages; their news consumption behavior is profoundly shaped by language choice. They often prefer local news and politics, sports, or health in their native language, but turn to a foreign language for domains such as technology, science, or entertainment, depending on content availability, topic-specific language preference, local relevance, trustworthiness, and internationality (Steichen et al., 2014; Ling et al., 2020). For example, a user might read local news in Croatian, international politics in English, and sports in German. Nevertheless, current state-of-the-art NRS only consider monolingual consumption patterns, and do not model language-dependent user preferences (Wu et al., 2023; Wang et al., 2025).

These user behaviors highlight why existing monolingual benchmarks and recommendation models are insufficient to address the complex cross-lingual information needs. Multilingual readers require additional effort when browsing news simultaneously in multiple languages, from switching between platforms, to manually synthesizing content across languages (Ling et al., 2020). As a result, they face more fragmented, less diverse, and less balanced news access (Ling et al., 2020), which reinforces discrepancies between monolingual and multilingual users, and further exacerbates the digital language divide in the news domain.

1.1.3 Challenges

The complexity and societal implications associated with developing responsible news recommendation technology demonstrate that we need to adopt a multifaceted approach for understanding, improving, and designing personalized neural recommendation methods. In this thesis, we aim to advance responsible and inclusive neural news recommendation by addressing the problem from several complementary angles: (i) analysis of common components, (ii) algorithmic bias and diversification, and (iii) multilinguality. Concretely, we systematically study the following challenges:

(C1) Analysis of Common Components: The effectiveness of a neural news recommender system depends on the interaction between its building blocks. Despite their interdependence, these components are often analyzed in isolation. *How do the design choices of common components – including architectural complexity, number of parameters, and computational demands – affect downstream performance, and can we improve over the commonly adopted set of components for more effective and efficient recommendation?*

(C2) Effects of Algorithmic Bias: Recommendation algorithms, typically optimized for relevance maximization, may implicitly favor certain types of content. *Which content-specific biases arise in neural news recommenders, and how do these biases correlate with user behavior?*

[7] https://ec.europa.eu/eurostat/statistics-explained/Foreign_language_skills_statistics
[8] https://southafrica-info.com/arts-culture/the-languages-of-south-africa/

(C3) Multi-Aspect Customization of Recommendations: Current beyond-relevance optimization efforts often target a single dimension, such as personalization or diversification, limiting practical applicability in scenarios where users require multiple criteria to be satisfied simultaneously. *Can we adapt neural news recommenders to efficiently support multi-aspect customization beyond pure relevance?*

(C4) Multilinguality: Language fundamentally shapes access to and engagement with personalized news delivery. *Can we foster inclusive neural news recommendation by accounting for multilinguality in both data and modeling approaches?*

For each of these challenges, we either conduct a series of experiments and extensive analyses and/or present new resources and/or approaches aimed at mitigating the identified issue. Through our findings, we hope to contribute to the development of more efficient, responsible, and inclusive news recommendation models that could be tailored to diverse user needs and surmount language barriers.

1.2 Contributions

This thesis constitutes a substantial contribution to the field of news recommendation by addressing the challenges associated with the multifaceted development of personalized neural news recommender systems. Drawing upon previously identified challenges, we focus both on improving the architectural design and diversity of NNRs and on enabling accurate recommendations irrespective of language boundaries. We now summarize our contributions, segmented into the following three topic areas, and highlight the key findings:

Analysis of Common Components. The performance of NNRs is determined by their foundational components: news and user encoders, scoring functions, and training objectives. Despite their importance, the field lacks a systematic understanding of how these components and their interplay drive performance. To address this gap and guide development of new models, we conduct two extensive analyses that decouple and evaluate these common components (**C1**). First, we explore alternative user modeling strategies and training objectives, identifying a simple yet highly effective approach for click behavior fusion. Subsequently, we systematically compare various encoder architectures across multiple criteria. We lastly synthesize these insights into concrete guidelines to aid researchers and practitioners make more informed decisions for model design.

1. *Alternatives for User Modeling and Training Objectives.* We are the first to introduce a unified framework to systematically and fairly compare NNRs across three crucial design dimensions: (i) candidate-awareness in user modeling, (ii) click behavior fusion, and (iii) training objectives (cf. Chapter 4). Within this framework, we challenge the standard approach of using complex user encoders (i.e., early fusion). We instead propose late fusion (LF), a simple and efficient method based on the mean-pooling of dot-product scores between candidate and clicked news embeddings. Our findings reveal that late fusion not only brings substantial performance gains across the board, but also reduces model complexity. Moreover, we demonstrate that supervised contrastive training is a viable alternative to standard point-wise classification objectives.

2. *Selection of Encoder Architectures.* Building on these findings, we move beyond one-sided, performance-only evaluations, which often shows negligible differences between models. We introduce a multifaceted evaluation protocol for news and user encoders that assesses: (i) representational similarity, (ii) recommendation overlap, and (iii) overall recommendation performance (cf. Chapter 5). This comprehensive analysis challenges the assumption that architectural complexity in neural news and user encoders is essential for accuracy. Instead, we find that performance hinges on the semantic accuracy of representations produced by news

encoders, while user encoders can be significantly simplified without sacrificing accuracy, a principle exemplified by the LF approach. Overall, this underscores the importance of understanding recommenders' behavior from multiple perspectives, and of balancing model complexity against performance when designing the architectures of NNR systems.

Multi-aspect Algorithmic Bias Analysis and Mitigation. We next investigate the effects of algorithmic news curation on users' selective exposure. Concretely, we analyze whether recommender systems of varying complexities exhibit biases towards specific news article aspects, such as sentiment, stance, or topics. (**C2**) To facilitate this, we compile two topic-specific corpora, in German and English, enriched with rich news metadata and user interaction data. Drawing upon the corresponding preliminary findings, we then propose a novel, modular framework for multi-aspect neural news recommendation that enables users to dynamically customize recommendations based on their preferences for either enhanced diversity or increased personalization (**C3**).

1. *Aspect-based Analysis of Bias in News Recommenders.* Our initial study assesses the sentiment and stance bias of four recommendation models using a self-compiled German news corpus on migration (cf. Chapter 6). Our findings reveal a slight tendency across all models to recommend articles with negative sentiments and stances. Notably, we identify a positive correlation between the bias of text-based recommenders and pre-existing user bias, suggesting a risk of amplifying readers' opinions and diminishing the diversity of recommended content. We next extend our analysis to NNRs and a corpus of news published in the US. To facilitate this, we augment our initial German corpus and construct a new dataset of English-language news on the same topic. Our investigation into the effects of bias and polarization with respect to topical category, sentiment, and political orientation reveals a trade-off between diversity and personalization. Specifically, we found that categorical and political personalization are more significantly aligned with content personalization than with the sentiment of news articles.

2. *Modular Multi-Aspect Neural News Recommendation.* To overcome the rigidity of conventional recommenders, which often "hardcode" aspect-based customization goals (e.g., diversification or personalization) directly into the model's architecture or training objectives, we propose MANNeR, a modular framework for multi-aspect neural news recommendation (cf. Chapter 7). MANNeR uses metric-based learning to train separate encoders for different news aspects (e.g., content, topic, sentiment). Its key innovation is the modular design that supports linearly combining the resulting aspect-specific similarity scores for final ranking. This approach yields two significant advantages at inference, without any model retraining: (i) creating ad-hoc custom ranking functions that reflect a user's preferences for each aspect, and (ii) dynamically identifying optimal trade-offs between content relevance and aspect-based customization. We demonstrate, through extensive experiments with topical categories and sentiment as additional aspects next to content itself, that MANNeR outperforms state-of-the-art NNRs across standard content-based recommendation, as well as single- and multi-aspect diversification and personalization tasks. Finally, we validate MANNeR's robustness, by showing that, when equipped with a multilingual PLM, its aspect encoder can be effectively transferred across languages.

Multilinguality in News Recommendation. In this part, we introduce the first multilingual, multi-parallel benchmark for analyzing the cross-lingual performance of neural news recommenders across diverse languages (**C4**). We then devise a domain-adaptation approach for sentence encoders that improves the cross-lingual transfer performance of NNRs, and propose a simple yet effective baseline for ZS-XLT in true cold-start and few-shot recommendation scenarios (**C4**).

1. *Benchmarking Cross-lingual News Recommendation.* We present xMIND, an open, multilingual news recommendation dataset for evaluating the cross-lingual transfer capabilities of

NNRs. Derived from the widely used English MIND dataset (Wu et al., 2020e) using machine translation (MT), xMIND covers 14 linguistically and geographically diverse languages, with digital footprints of varying sizes (cf. Chapter 8). The multi-parallel nature of our benchmark, where the same news articles and user histories are present in all 14 languages, enables the direct comparison of model behavior on both high- and low-resource languages, as well as the simulation of multilingual news consumption by polyglot readers. We use xMIND to systematically evaluate various content-based NNRs in ZS-XLT and FS-XLT settings, examining both monolingual and bilingual consumption patterns. We find that current NNRs, even when grounded in multilingual language models, suffer significant performance drops in ZS-XLT. Moreover, our results reveal that in FS-XLT scenarios, the inclusion of target-language data in training provides surprisingly limited benefits, particularly for bilingual news consumers. Our analyses highlight that existing NNRs still struggle to provide accurate recommendation to users speaking resource-scarce languages, particularly those linguistically distant from the source language (typically English).

2. *Domain Adaptation for Cross-lingual Recommendation.* As a solution to the aforementioned issues, we propose training a news-adapted sentence encoder (NaSE). Concretely, we domain-specialize a pretrained, massively multilingual sentence encoder with denoising auto-encoding and MT objectives on PolyNews and PolyNewsParallel, two multilingual, news-specific corpora that we compiled for this purpose (cf. Chapter 9). We empirically show that NaSE yields consistent performance gains over general-purpose LMs in ZS-XLT, especially when its parameters are frozen during downstream fine-tuning. These findings challenge the established paradigm of fine-tuning LMs on user click data, a strategy that is often infeasible in data-scarce (e.g., no news data in low-resource languages) or cold-start contexts (e.g., missing click logs for new users). Building on this, we establish a simple and strong baseline based on frozen NaSE embeddings and late click behavior fusion. This approach achieves state-of-the-art performance in ZS-XLT in true cold-start and few-shot news recommendation.

An overview of all the resources published in this thesis is available in Appendix A. We hope that our work inspires future research toward the development of more customizable, diverse, and inclusive neural news recommendation technology.

1.3 Outline

The thesis is organized into five parts, each comprising multiple Chapters.

In *Part I Fundamentals*, we establish the theoretical foundations of this thesis, covering fundamental concepts and terminology in representation learning (Chapter 2) and neural news recommender systems (Chapter 3). This part also details the evaluation resources, including experimental protocol, metrics, datasets, and frameworks, relevant to our experiments (Section 3.2-Section 3.3).

In *Part II Analysis of Common Components*, we analyze architectural design choices of NNRs. First, in Chapter 4, we investigate alternative user modeling methods and training objectives to improve the efficiency and accuracy of NNRs. In Chapter 5, we present an in-depth study of representative news and user encoders, proposing key strategies for optimal model design (**C1**).

In the subsequent *Part III Multi-Aspect Bias Analysis and Mitigation*, we construct multiple topic-specific news and user corpora and show that NRS are prone to biases regarding certain news aspects, such as sentiment, stance, or political leaning (**C2**, Chapter 6). Drawing upon these insights, in Chapter 7 we introduce a modular framework for multi-aspect neural news recommendation which leverages metric-based learning to enable on-the-fly customization (i.e., diversification or personalization) of recommendations (**C3**).

In *Part IV Multilinguality*, we describe our efforts to systematically address the multilinguality challenge (**C4**). To this end, we use machine translation to construct a multilingual multi-parallel

news recommendation dataset from the MIND benchmark, spanning geographically and linguistically diverse languages. We then perform multi-layered analyses, revealing that existing NNRs, even when based on multilingual LMs, suffer drastic performance losses in target languages in both ZS-XLT and FS-XLT compared to source-language performance (Chapter 8). Consequently, in Chapter 9, we propose to adapt a massively multilingual sentence encoder to the news domain. We further demonstrate that using this adapted sentence encoder as the recommender's news encoder achieves stronger ZS-XLT performance, particularly in true cold-start and few-shot recommendation scenarios.

Lastly, *Part V* summarizes the main findings and contributions, and outlines potential directions for future work toward more effective, responsible, and inclusive recommendation (Chapter 10).

Part I

FUNDAMENTALS

CHAPTER 2

REPRESENTATION LEARNING

The central task of a natural science is to make the wonderful commonplace: to show that complexity, correctly viewed, is only a mask for simplicity; to find pattern hidden in apparent chaos.

HERBERT A. SIMON
«THE SCIENCES OF THE ARTIFICIAL »

Neural news recommendation models rely on effective news modeling to understand and represent the content of articles. News modeling, in turn, requires transforming raw textual data into numerical representations that can be processed by computational methods. In this Chapter, we review representation learning approaches for modeling the semantic content of news articles. We begin in Section 2.1 by tracing the evolution of language representation learning, from statistical and early neural language models to static word embeddings and contextualized representations with Transformers. We then discuss the pretraining of language models in Section 2.2, and conclude with a review of multilingual representation learning in Section 2.3. This Chapter lays the foundational background for the neural news recommendation models discussed in the remainder of the thesis.

2.1 Language Modeling

Language modeling is a foundational task in natural language processing (NLP), concerned with estimating the probability of sequences of tokens, typically words or characters, in a language. Formally, a language model (LM) is a probabilistic model that assigns a likelihood to each token given its preceding context (Jurafsky and Martin, 2025). Modern large-scale LMs attain such high accuracy in modeling natural language that they can be applied to a variety of downstream tasks in a generative fashion, including summarization, question answering, and text completion.

Early approaches to language modeling were predominantly based on statistical methods, most notably n-gram models. These models estimate the probability of a word given the preceding $n-1$ tokens by computing frequency counts of n-grams (i.e., sequences of n tokens) observed in a training corpus (Jurafsky and Martin, 2025). Despite their simplicity and interpretability, statistical language models are ill-equipped to handle the complexity and ambiguity of natural language, due to data sparsity and inability to capture long-range dependencies. The number of possible token sequences grows exponentially with vocabulary size; for instance, for a vocabulary V, there exist $|V|^{10}$ possible sequences of length ten, a majority of which could be valid but not included in the training data. Consequently, statistical language models rely solely on local co-occurrence statistics observed within limited context windows during training, resulting in poor generalization to rare or unseen sequences.

Bengio et al. (2003) hence introduced the *neural language model*, which leverages distributed representations of words learned through neural networks to address this "curse of dimensionality". Their approach employed a shallow feed-forward architecture to map discrete tokens into dense,

low-dimensional vectors, commonly referred to as *word embeddings*, by predicting the next token in a sequence. By embedding words in a shared continuous space $\mathbb{R}^d$, where $d \ll |V|$, neural language models substantially improve generalization to similar but unseen contexts, while dramatically reducing the number of model parameters. This paradigm shift laid the foundation for widely adopted static word embedding methods such as word2vec (Mikolov et al., 2013) or fastText (Bojanowski et al., 2017), which learn word representations from large corpora based on distributional similarity and surrounding context. However, static embeddings assign a single vector to each word, irrespective of context, and are constrained by fixed context windows. As a result, they fail to capture contextual nuances, struggle with semantic ambiguity (e.g., polysemy), and cannot adequately handle out-of-vocabulary (OOV) words.

These shortcomings motivated the development of contextualized representations produced by transformer-based LMs, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b). Such models employ subword unit tokenization schemes (Gage, 1994; Sennrich et al., 2016) – a property already present in models like fastText (Bojanowski et al., 2017) – to alleviate OOV issues, while leveraging self-attention mechanisms to compute context-dependent representations for each token. This architectural development enables modeling of long-range dependencies and yields substantial performance improvements across a broad spectrum of downstream NLP tasks. In the following sections, we delve deeper into the key concepts and milestones that have shaped the evolution language models.

2.1.1 From Statistical to Neural Language Models

Formally, LMs define a joint probability distribution $P(t_1, ..., t_N)$ over a token sequence S consisting of N tokens $[t_1, ..., t_N]$. As such, an LM can assign a probability to any word sequence or sentence. Given a sequence of tokens $[t_1, t_2, ..., t_N]$, an LM computes the probability of the entire sequence (i.e., the likelihood that the tokens occur jointly) by applying the chain rule of probability (Jurafsky and Martin, 2025):

$$P(t_{1:N}) = P(t_1)P(t_2|t_1)...P(t_N|t_{1:N-1}) = \prod_{i=1}^{N} P(t_i|t_{1:i-1}) \tag{2.1}$$

Language modeling can therefore be viewed as learning a sequence of conditional probability distributions, each predicting the next token given its preceding context. However, in practice, an overwhelming number of possible token sequences are never observed in the training corpus, making it infeasible for statistical LMs to reliably estimate these conditional probabilities directly. To address this challenge, statistical LMs adopt the Markov assumption, which posits that the probability of a token depends only on a limited number of preceding tokens rather than the entire history (Jurafsky and Martin, 2025). Under this assumption, the conditional probability of a token is approximated by conditioning only on the previous $n-1$ tokens, or n-grams:

$$P(t_N|t_1, ..., t_{N-1}) \approx P(t_N|t_{N-n+1:N-1}) \tag{2.2}$$

This approximation exploits the intuition that tokens closer to each other in a sequence tend to exhibit stronger statistical dependence (Bengio et al., 2003). Despite the simplicity, statistical LMs like n-gram suffer from several fundamental limitations (Bengio et al., 2003). Most notably, their reliance on the Markov assumption restricts the effective context to a short, fixed window, severely limiting their ability to capture long-range dependencies. As a result, these models often fail to represent broader contextual information and may produce repetitive patterns. Furthermore, n-gram LMs treat words as discrete symbols and rely exclusively on frequency counts, thereby ignoring syntactic and semantic similarities between words. For example, words such as “cat” and “dog” frequently occur in similar contexts and should thus provide similar contextual cues when predicting surrounding words.

However, because n-gram models treat words as discrete symbols and rely solely on frequency counts, they cannot exploit such semantic relatedness across contexts. Finally, statistical LMs struggle with words that are not present in the training vocabulary, leading to OOV issues.

These limitations have led to the development of more advanced language modeling approaches that relax the Markov assumption through architectural innovations, ultimately enabling improved generalization to unseen contexts and more effective modeling of natural language.

2.1.2 Static Word Embedding Models

Dense word representation models project each token t (typically corresponding to a single word) from a vocabulary V into a low-dimensional latent semantic space $\mathbb{R}^d$, where $d \ll |V|$. In static embedding spaces, each word is associated with a single, fixed vector representation that does not vary with the context in which the word appears. Consequently, a polysemous word such as "court" receives the same embedding whether it occurs in "The witnesses testified today in *court* in the famous lawsuit.", "Some teenagers were playing football on the basketball *court*.", or "The politician tried to *court* favor with voters by promising new projects.".

The training objectives of word embedding models are grounded in the distributional hypothesis (Harris, 1954), often summarized by the principle that "a word is characterized by the company it keeps" (Firth, 1957). Under this assumption, words that occur in similar contexts are expected to share semantic meaning. Accordingly, word embedding models aim to place words that appear in similar contexts closer together in the representation space $\mathbb{R}^d$, such that semantic similarity is reflected by geometric proximity. The resulting distributed word representations have been shown to be highly effective across a wide range of downstream NLP tasks, such as lexical analogical reasoning or text classification (Mikolov et al., 2013; Arora et al., 2017). In the following paragraphs, we review prominent algorithms for learning static word embeddings, including word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017).

word2vec. With word2vec, Mikolov et al. (2013) introduced two widely used word representation models based on a shallow feed-forward neural network architecture: the Skip-Gram (SG) and the Continuous Bag-of-Words (CBOW) models. These two variants differ in both their architectural design and training objectives. The SG model predicts surrounding context words given a target (center) word, whereas the CBOW model predicts the target word from its surrounding context.

Both models are trained on large-scale text corpora using a sliding context window. For a center word w_t, its context is defined as $\mathcal{C}(w_t) = \{w_{t-n}, ..., w_{t-1}, w_{t+1}...., w_{t+n}\}$, consisting of n neighboring words to the left and to the right of w_t, respectively. Each unique term $w_i \in V$ in the vocabulary V is associated with two d-dimensional vector representations: a center embedding $\mathbf{e_{w_i}}$ and a context embedding $\mathbf{u_{w_i}}$. These correspond to the i^{th} rows of the center embedding matrix $\mathbf{E} \in \mathbb{R}^{|V| \times d}$ and the context embedding matrix $\mathbf{U} \in \mathbb{R}^{|V| \times d}$, respectively. The training objective encourages words that co-occur within the same context window to have similar vector representations by maximizing the similarity between the embeddings of center and context words.

Skip-Gram. Given a sequence of words $[w_{t-n}, ..., w_t, ..., w_{t+n}]$ with the center word w_t, the goal of the SG model is to predict the surrounding context words $\mathcal{C}(w_t)$. The model estimates the conditional probability of observing a context word $w_c \in \mathcal{C}(w_t)$ given the center word w_t using softmax:

$$P(w_c|w_t) = \frac{\exp(\mathbf{e_{w_t}^T u_{w_c}})}{\sum_{j=1}^{|V|} \exp(\mathbf{e_{w_t}^T u_{w_j}})} \tag{2.3}$$

Computing the normalization term in the denominator of Eq. (2.3) is computationally expensive for large vocabularies with millions of words (Goldberg and Levy, 2014). Therefore, Mikolov et al. (2013) proposed *negative sampling* as an efficient approximation to the full softmax computation. Under this formulation, instead of estimating probabilities over the entire vocabulary V in the training

corpus, the model only predicts probabilities over a subset of the vocabulary, consisting of the true context word and k randomly sampled negative examples.

The SG model is trained by minimizing the negative log-likelihood over all observed center-context pairs in the corpus:

$$L_{SG} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{w_c \in \mathcal{C}(w_t)} \log P(w_c|w_t) \tag{2.4}$$

where T denotes the total number of terms in the sequence.

Continuous Bag-of-Words. In contrast to SG, the CBOW model predicts the center word w_t given its context $\mathcal{C}(w_t)$. The model is trained by minimizing the negative log-likelihood of the center word conditioned on its context:

$$L_{CBOW} = -\log P(w_t|\mathcal{C}(w_t)) \tag{2.5}$$

where $P(w_t|\mathcal{C}(w_t))$ denotes the probability of observing w_t as the center word given its surrounding context $\mathcal{C}(w_t)$.

GloVe. Another widely used static word embedding model is GloVe proposed by Pennington et al. (2014). In contrast to word2vec, which learns dense representations through local context predictions, GloVe derives word embeddings from global corpus-wide co-occurrence statistics. Rather than relying on direct co-occurrence probabilities, GloVe learns word representations by optimizing an objective function that implicitly factorizes a word-word co-occurrence matrix. The key intuition underlying this approach is that the ratios of co-occurrence probabilities are more informative of semantic relationships than absolute co-occurrence frequencies. Specifically, given two words w_i and w_j, and a probe word w_k, the ratio $\frac{P(w_i|w_k)}{P(w_j|w_k)}$ captures how strongly w_k is associated with w_i relative to w_j. By learning embeddings that preserve these probability ratios across the corpus, GloVe effectively encodes meaningful semantic patterns while leveraging global statistical information.

fastText. Traditional word embedding models, such as word2vec, assign a single vector representations to each word, disregarding its internal morphological structure. As a result, these models struggle to represent rare and OOV words. Recognizing these limitations, Bojanowski et al. (2017) introduced fastText, an extension of the SG model that incorporates subword information to learn richer lexical representations. In fastText, each word is represented as a collections of its constituent character n-grams, with each character n-gram associated with a distinct embedding.[1] The dense vector representation of a word is obtained by summing the embeddings of its character n-grams, which are learned using a SG model. This formulation enables the model to generate meaningful representations for rare or unseen words by composing them from shared subword units. Moreover, by frequently updating shared character n-gram representations across multiple words, fastText yields more robust embeddings for morphologically rich languages.

2.1.3 Contextualized Representations

Word representation models such as word2vec, GloVe, or fastText suffer from two fundamental limitations: (i) they produce context-agnostic representations, and (ii) they are unable to effectively capture long-range linguistic dependencies. First, the embeddings learned by these models are static and context-insensitive, as they collapse multiple senses of a word into a single vector that reflects a frequency-weighted average of its usages in the training corpus (Arora et al., 2018). Consequently, such representations cannot adequately model polysemy or contextual meaning variations. Second, the training objectives of static word representation models define a word's context using a narrow window of neighboring tokens. This local view of context prevents them from capturing long-range

[1] Typical character n-gram lengths are in the range 3 to 6.

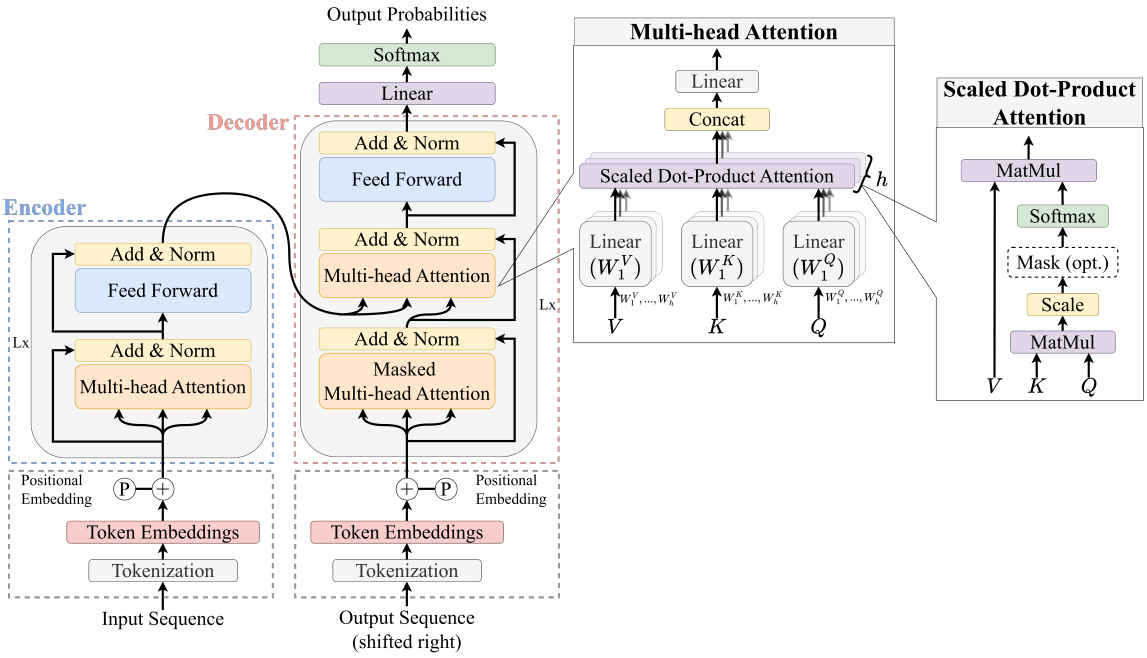


Figure 2.1: Overview of the Transformer architecture introduced by Vaswani et al. (2017). The Encoder (left) and Decoder (center left) each consist of L stacked layers. Multi-head Attention (center right) consists of several attention layers running in parallel, allowing the model to capture different aspects of the input by attending information from diverse representation subspaces at different positions. $\mathbf{V}, \mathbf{K}$, and $\mathbf{Q}$ denote the learnable value, key, and query matrices, whereas $\mathbf{W}_i^{\mathbf{V}}$, $\mathbf{W}_i^{\mathbf{K}}$, and $\mathbf{W}_i^{\mathbf{Q}}$ denote trainable weight matrices. The Scaled Dot-Product Attention (right) computes contextualized token representations by modeling dependencies between tokens. The figure is adapted from Vaswani et al. (2017).

dependencies that are intrinsic to natural language, where relevant information may span entire sentences or paragraphs.

As a result of these shortcomings, the paradigm of pretraining static word embeddings was largely superseded by deep LMs that generate dynamic, context-dependent token representations. These models mitigate the limitations of static embeddings and have led to substantial performance gains in downstream tasks. Early contextualized LMs, most notably ULMFit (Howard and Ruder, 2018) and ELMO (Peters et al., 2018), employed variants of long short-term memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) as sequence encoders. However, recurrent neural network-based LMs struggled to model long-range dependencies and were inherently difficult to parallelize due to their sequential nature. In the following, we focus on Transformer-based LMs (Vaswani et al., 2017), which addressed these limitations and fundamentally reshaped NLP. Transformers have since become the state-of-the-art architecture for learning text representations and underpin modern large language models (LLMs), including Llama (Touvron et al., 2023) and GPT (OpenAI et al., 2024).

Figure Fig. 2.1 illustrates the original Transformer architecture introduced by Vaswani et al. (2017). We next provide a step-by-step discussion of how input sequences are processed, the core building blocks of the Transformer, and its language modeling head.

Subword Tokenization. In parallel with the architectural progression of Transformer-based models, subword tokenization emerged as a crucial advance in textual preprocessing and has since been adopted by the majority of Transformer-based LMs, which decompose input text into subword units (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020, *inter alia*). Formally, a tokenizer T is an algorithm that decomposes a text sequence $S_{i=1}^{|D|}$ from the training corpus $|D|$ into

a sequence of tokens $\bigcup_{i=1}^{|D|} T(S_i)$, which serve as input to the LM. To this end, the tokenizer first constructs a vocabulary $V = \{t_1, ..., t_{|V|}\}$ from the training corpus D. By sharing the subword representations across multiple words, subword tokenization reduces the number of parameters required to encode lexical units, thereby decreasing the overall model size. Moreover, it enables models to represent a substantially larger effective vocabulary and to obtain meaningful representations for rare words or morphological variants that were not not observed during training.

Static word representation models (cf. § 2.1.2), relied on rule-based tokenization schemes that first segmented text into sentences using punctuation and subsequently split sentences into tokens based on whitespace. This approach treats tokens as atomic input units and ignores the internal morphological structure of words. As a consequence, the vocabulary expands with every new term encountered, often leading to excessively large vocabularies. Despite this growth, such LMs still handle OOV tokens poorly at test time (Sennrich et al., 2016). This limitation is particularly pronounced in morphologically rich and agglutinative languages, which form complex words by combining multiple morphemes, each conveying a distinct grammatical meaning. For example, in Finnish, the word *talo|ssa|ni* ("in my house") is derived from the root morpheme *talo* ("house") combined with the case suffix *-ssa* ("in") and the possessive suffix *-ni* ("my"). A word embedding model that has not encountered *talo|ssa|ni* during training would be unable to represent it meaningfully. In contrast, a Transformer-based LM employing subword tokenization can often mitigate the OOV problem by splitting rare words into known subword units such as *talo* and shorter frequent character sequences. However, these subword units do not necessarily correspond to true morphemes, and their segmentation is determined by frequency statistics rather than linguistic boundaries. While dedicated morphologically driven tokenization methods have been proposed to better capture such structure, these require a language-specific morphological analyzer for each language (Park et al., 2021; Hofmann et al., 2021; García et al., 2025; Asgari et al., 2025).

To address the limitations of word-level tokenization, more advanced subword tokenization techniques have been proposed, which are typically trained on large corpora prior to LM pretraining (Sennrich et al., 2016; Wu et al., 2016; Kudo and Richardson, 2018; Kudo, 2018, *inter alia*). Among these, Byte Pair Encoding (BPE) (Sennrich et al., 2016) and WordPiece (Wu et al., 2016) have emerged as two of the most widely adopted tokenization algorithms. Both approaches decompose input text into variable-length character sequences drawn from a vocabulary V learned during training. After an initial whitespace tokenization step, the vocabulary is initialized with all unique characters observed in the corpus. The vocabulary is then iteratively expanded: BPE merges the most frequent pairs of adjacent tokens, whereas WordPiece adds subword units that maximize the likelihood of the training data under a language modeling objective. This iterative process continues until a user-specified vocabulary size is reached. At inference time, input text is first segmented into characters and then successively merged into subword units according to either learned merge operations in the case of BPE or a greedy longest-match first strategy in the case of WordPiece. As a result, frequent words are more likely to be represented as single tokens, while rare or unseen words are decomposed into multiple subword units.

By construction, BPE and WordPiece alleviate issues associated with closed vocabularies and explicitly account for morphological variation. However, subword tokenizers face certain deficiencies. They are prone to over-segmentation when confronted with noisy inputs, typological errors, or domain shifts, which can adversely affect downstream performance (Kumar et al., 2020; Sun et al., 2023). Consequently, the robustness and generalization of Transformer-based LMs can be sensitive to the quality and suitability of the tokenization method. In the news domain, such shortcomings are especially relevant due to the prevalence of emerging named entities (e.g., people, organizations, events), domain-specific terminology, spelling variations, and multilingual content, which increase the likelihood of over-segmentation and incosistent token representations.

In Transformer-based LMs, a subword tokenizer with vocabulary V first splits an input into N subword tokens $t_1, ..., t_N$. We next describe how these tokens are processed and fed into the LM.

Input Embeddings. Transformer-based LMs typically operate on two types of input embeddings: token embeddings and positional embeddings. Token embeddings represent individual tokens as dense vectors that capture their lexical and semantic properties and are learned during pretraining. These embedding may subsequently be fine-tuned to adapt the model to specific downstream tasks.

Unlike recurrent architectures, transformers have no inherent notion of token order, as the underlying self-attention mechanism is permutation-invariant. To inject information about token position, positional embeddings are combined with token embeddings. This positional information enables the model to account for word order and distinguish between sequences with identical tokens but different arrangements, such as "bear bites man" and "man bites bear". The original Transformer architecture introduced by Vaswani et al. (2017) employed fixed sinusoidal positional embeddings, which encode relative position information implicitly. However, these positional embeddings are not learned and therefore cannot be easily adapted to new tasks or extended to longer input sequences after LM pretraining (Chen et al., 2023b). Consequently, several alternatives have been proposed, including relative positional embeddings (Shaw et al., 2018), rotary positional embeddings (Su et al., 2024) or bias-based positional encoding schemes (Press et al., 2022), each designed to address limitations of absolute positional embeddings.

Formally, given a tokenized input sequence of length N, $[t_1, ..., t_N]$ with $t_i \in V$, each token t_i is mapped to a vector $\mathbf{e_{t_i}}$ using a token embedding matrix $\mathbf{E} \in \mathbb{R}^{|V| \times d}$. In parallel, a positional embedding $\mathbf{p_i} \in \mathbb{R}^d$ is assigned to the i^{th} position in the sequence. The initial hidden representation of token t_i is obtained by the element-wise addition of its token and positional embeddings: $\mathbf{h_{t_i}^{(0)}} = \mathbf{e_{t_i}} + \mathbf{p_i}$. The resulting sequence of vertically stacked embeddings $[\mathbf{h_1^{(0)}}, ..., \mathbf{h_N^{(0)}}] \in \mathbb{R}^{|V| \times d}$ is then passed through the L transfomer layers.

Transformer Layers. The Transformer was introduced by Vaswani et al. (2017) as a sequence-to-sequence model originally designed for machine-translation. It employs a multi-layer encoder-decoder architecture, as depicted in Fig. 2.1. The encoder is composed of L identical layers, which are recursively applied to transform an input sequence of N subword tokens, $[t_1, ..., t_N]$, into a sequence of contextualized token embeddings $[\mathbf{h_1^{(L)}}, ..., \mathbf{h_N^{(L)}}] \in \mathbb{R}^{N \times d}$. Each Transformer layer L_i consists of a multi-head self-attention module and a feed-forward sub-network (FFN), combined with layer normalization and residual connections. The decoder has a similar structure, comprising L identical layers. However, it additionally integrates the encoder outputs via a cross-attention module and employs masked multi-head self-attention to prevent inputs from attending to future positions. This masking ensures that each decoding step only considers information available from preceding positions, which is crucial for autoregressive generation. In the following, we describe the core building blocks of the encoder, namely the *Transformer layers*.

Multi-Head Self-Attention. Vaswani et al. (2017) introduced the *self-attention* mechanism to compute the relevance of each token in a sequence based on its contextual relationship with other tokens. Self-attention enables the model to capture dependencies between tokens, regardless of their distance in the sequence, and is based on the scaled dot-product attention. First, the input hidden representation are independently projected into three learnable matrices: the query matrix $\mathbf{Q} \in \mathbb{R}^{d_k}$, the key matrix $\mathbf{K} \in \mathbb{R}^{d_k}$, and the value matrix $\mathbf{V} \in \mathbb{R}^{d_v}$. Given an input matrix $\mathbf{X} \in \mathbb{R}^d$, where each row $\mathbf{x}_i$ corresponds to a token embedding, the projections is computed as:

$$\mathbf{Q} = \mathbf{X}\mathbf{W^Q}; \mathbf{K} = \mathbf{X}\mathbf{W^K}; \mathbf{V} = \mathbf{X}\mathbf{W^V} \tag{2.6}$$

where $\mathbf{W^Q}, \mathbf{W^K} \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W^V} \in \mathbb{R}^{d \times d_v}$ are trainable weight matrices. Each row of the resulting matrices corresponds to the query $\mathbf{q}_i$, key $\mathbf{k}_i$, and value $\mathbf{v}_i$ vectors for the i^{th} token.

The attention mechanism computes the relevance of each token to every other token by measuring the similarity between each query and all the keys, scaling by the factor $\sqrt{d_k}$ to stabilize gradients, and applying a softmax to obtain a probability distribution over the sequence. The resulting attention scores weight the value vectors to produce context-aware token representations:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2.7}$$

Instead of using a single self-attention mechanism, *multi-head self-attention* uses h attention heads, each with independent projection matrices. This allows the model to attend to multiple aspects of the input simultaneously and capture diverse types of dependencies. Prior work has shown that different attention heads learn to capture both general patterns, by attending over the entire sequence, the punctuation, or a specific token, as well as syntactical relationships, including coreference or predicate-object dependencies (Clark et al., 2019). Each attention head is computed as:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}) \tag{2.8}$$

where $\mathbf{W}_\mathbf{i}^\mathbf{Q}, \mathbf{W}_\mathbf{i}^\mathbf{K} \in \mathbb{R}^{d\times d_k}$, and $\mathbf{W}_\mathbf{i}^\mathbf{V} \in \mathbb{R}^{d\times d_v}$ are the learnable projection matrices for the i^{th} head. Attention computations for all heads are performed in parallel, improving computational efficiency. The outputs of all h heads are then concatenated and projected back to the model dimensionality d using a linear transformation $\mathbf{W}^O \in \mathbb{R}^{hd_v\times d}$:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}^O \tag{2.9}$$

The number of attention heads is a key hyperparmeter that determines the model's capacity to capture different patterns and dependencies within the sequence, with more heads enabling finer-grained representations at the cost of increased computational complexity.

Feed-Forward Sub-Network. After the attention sub-layer, a *position-wise fully connected FFN*, consisting of a two-layer neural network with a ReLU activation (Glorot et al., 2011), is applied as:

$$FFN(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{2.10}$$

where the same FFN parameters are applied independently to each token position. Equivalently, the FFN can be interpreted as two successive position-independent convolutions with kernel size 1. Prior work has shown that FFNs encode a substantial amount of factual and linguistic knowledge acquired by the Transformer during pretraining (Geva et al., 2021; Meng et al., 2022).

Normalization and Residual Connections. Each attention and FFN is wrapped with a *residual connection* (He et al., 2016), which adds the sub-layer's input to its output. Residual connections help preserve and propagate useful information across layers, thus mitigating the vanishing gradient problem.[2] Following each residual connection, *layer normalization* (Ba et al., 2016) is applied to the resulting representations. This normalizes each token representations across the feature dimension d, effectively standardizing the scale of hidden representations, stabilizing gradients, and accelerating training. Together, residual connections and layer normalization significantly enhance the training stability and efficiency of deep neural networks.

Language Modeling Head. The Transformer's decoder outputs a sequence of hidden representations, each of which is passed through a linear projection layer followed by a softmax function to produce a probability distribution over a fixed vocabulary. In the original architecture proposed by Vaswani et al. (2017), the weights of this output projection layer are tied to the input token embedding matrix. In practice, transformers are often extended with task-specific output heads, which typically consist of one or more linear layers followed by a softmax or other suitable activation function. These heads map the learned representations to the task-dependent output space, such as class labels in text classification or vocabulary tokens in sequence generation.

[2]The vanishing gradient problem arises in deep neural networks when gradients diminish exponentially as they are backpropagated through many layers, eventually approaching zero. As a result, early layers receive little to no learning signal, which slows convergence and hampers effective training. Residual connections alleviate this issue by providing shortcut paths that allow gradients to flow more directly through the network.

2.2 Pretrained Language Models

2.2.1 Language Model Pretraining

The Transformer architecture enables LMs to effectively model contextual information in input sequences, capture long-range dependencies, and process data efficiently through parallel training. Its modularity and adaptability has driven the success of Transfomer-based architectures and has led to the emergence of Pretrained Language Models (PLMs), marking a major breakthrough in NLP and yielding substantial performance gains across a wide range of tasks.[3] Most LMs can be broadly categorized into encoder-only models (Devlin et al., 2019; Conneau and Lample, 2019; Liu et al., 2019b; Warner et al., 2025, *inter alia*), which are particularly effective for natural language understanding tasks (e.g., text or token classification, sentence similarity, information retrieval) and decoder-only models (Yang et al., 2019b; Xue et al., 2021; Touvron et al., 2023; OpenAI et al., 2024; DeepSeek-AI et al., 2024, *inter alia*), which excel at generative tasks (e.g., text generation, summarization, translation) due to their autoregressive nature.

PLMs are typically trained in a two-stage process. First, during *pretraining*, the model is trained on massive (often web-scale) unlabeled text corpora drawn from heterogeneous sources, using language modeling objectives. In this phase, the models learns statistical regularities and contextual representations of the words, thereby encoding linguistic and factual knowledge (Petroni et al., 2019; Geva et al., 2021; Dai et al., 2022). Second, in the *fine-tuning* stage, the pretrained LM is further trained on labeled data for specific downstream tasks, such as text classification or sentence similarity.

In what follows, we focus exclusively on encoder-only Transformer-based PLMs, as these are the most relevant to recommendation applications (see Section 3.1). We next introduce two prominent encoder-based models: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b).

Bidirectional Encoder Representations from Transformers (BERT). BERT, proposed by Devlin et al. (2019), is the first PLM to build upon the Transformer architecture by employing a multi-layer bidirectional Transformer encoder, as described in Section 2.1. By modeling both left and right context simultaneously, BERT learns deeply contextualized token representations that are well-suited for a wide range of natural language understanding tasks.

BERT uses a WordPiece tokenizer (Wu et al., 2016) to segment input text into subword units. Because the model is pretrained on sentence pairs, it incorporates segment embeddings to indicate which sentence each token belongs to. Additionally, BERT introduces several special tokens: a separator token `[SEP]` to mark sentence boundaries, a padding token `[PAD]` to fill unused positions in variable-length input, and a classification token `[CLS]` that enables the model to learn a sequence-level representation for downstream tasks. These special tokens are embedded in the same space as the vocabulary tokens. Each input token is represented as the sum of its token embedding, position embedding, and segment embedding. For a sentence pair, the input sequence is formatted as `[CLS]` $t_{11}...t_{1i}$ `[SEP]` $t_{21}...t_{2j}$ `[SEP]`. BERT is pretrained on a large corpus consisting of the English Wikipedia and the BookCorpus (Zhu et al., 2015) using two self-supervised learning objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), depicted in Figure Fig. 2.2.

The MLM objective, often referred to as a *Cloze* task (Taylor, 1953), is a token-level prediction task in which the input sequence is partially corrupted and the model is trained to recover the original tokens. Specifically, 15% of the tokens in an input sequence $[t_1, ..., t_N]$ are selected for prediction. Of these, 80% are replaced with a special `[MASK]` token, 10% are replaced with a randomly sampled token from the vocabulary V, and the remaining 10% are left unchanged. The model is trained to predict the original tokens based on their surrounding context $\{t_1, ..., t_N\} \setminus t_i$, enabling the learning of bidirectional contextual dependencies at the token level.

In contrast, NSP is a sentence-level binary classification task designed to capture relationships

[3] For brevity, throughout the remainder of this thesis, we use the term *pretrained language models (PLMs)* to refer to all pretrained Transformer-based LMs.

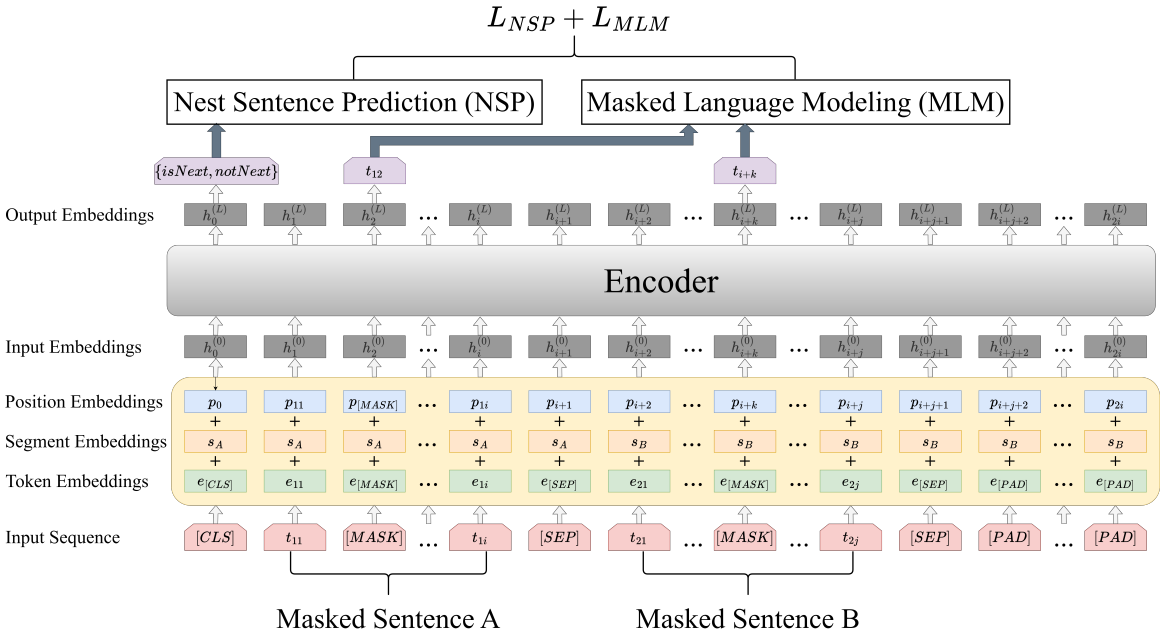


Figure 2.2: Overview of the BERT pretraining architecture. A training instance is formed by concatenating two text sequences, `A` containing i subword tokens and `B` containing j subword tokens, augmented with a special classification token `[CLS]`, sequence separator tokens `[SEP]`, and padding tokens `[PAD]` as needed. The resulting input sequence is partially corrupted by randomly masking a subset of tokens using the `[MASK]` symbol. Each input embedding $h_i^{(0)}$ to the encoder is computed as the sum of the token, segment, and positional embeddings. These embeddings are then contextualized by a Transformer encoder. BERT is pretrained using two self-supervised objectives: Masked Language Modeling (MLM), which reconstructs the masked tokens from their contextualized representations $h_i^{(L)}$ in the final layer, and Next Sentence Prediction (NSP), which predicts, using the final hidden representation of the `[CLS]` token $h_{[CLS]}^{(L)}$, whether sentence `B` follows sentence `A` in the training corpus or is randomly sampled. The figure is adapted from Devlin et al. (2019).

between sentences and improve sequence-level representations. The model is trained to predict whether two concatenated input sentences `A` and `B` are consecutive in the original corpus or randomly paired (with equal probability). The final hidden representation of the `[CLS]` token is passed to a classification head with a binary softmax to make this prediction. Both the MLM and the NSP objectives are optimized jointly during pretraining.

Robustly Optimized BERT Approach (RoBERTa). In subsequent work, Liu et al. (2019b) observed that the original BERT model was undertrained and proposed an improved training strategy, resulting in the RoBERTa model. While RoBERTa retains the same Transformer architecture as BERT, it introduces several key improvements in the training procedure. First, RoBERTa is pretrained using a significantly larger batch size (up to 8K, compared to 256 in BERT) and on substantially larger and more diverse text corpora, approximately an order of magnitude more data than used for BERT. Second, the NSP objective is removed, and the original MLM task is enhanced by dynamically masking tokens during training. Instead of applying a single static masking pattern during preprocessing, RoBERTa resamples masked positions at each training epoch, exposing the model to a broader range of prediction contexts. These modifications to the training data and objectives lead to consistent performance improvements over BERT across a wide range of downstream tsks, demonstrating the importance of pretraining scale and objective design in Transformer-based LMs.

2.2.2 Sequence-Level Representations

Although token-level representations are sufficient for certain tasks, such as token classification or named entity recognition, learning a single dense representation for long-range sequences, such as sentences or entire documents, is crucial for tasks like text classification, semantic textual similarity, and retrieval (Reimers and Gurevych, 2019; Muennighoff et al., 2023).

The emergece of static word embedding models (cf. § 2.1.2) motivated subsequent research on end-to-end sequence representation learning. Notably, doc2vec (Le and Mikolov, 2014) extended the SG and CBOW objectives to learn paragraph- or document-level embeddings by prepending paragraph-specific vectors in the language modeling objectives. However, doc2vec faced several limitations, including sensitivity to document length and training hyperparameters (Lau and Baldwin, 2016), as well as the need for costly retraining to handle unseen sequences. Consequently, approaches that aggregate static word embeddings into sequence-level representations – typically through weighted averages of token embeddings – were preferred, achieving competitive performance on tasks like text classification or retrieval while maintaining computational efficiency (Arora et al., 2017; Schmidt et al., 2019).

The introduction of the Transformer architecture and PLMs led to the development of Transformer-based sentence encoders (SEs). A prominent example is S-BERT (Reimers and Gurevych, 2019), which demonstrated that BERT or RoBERTa can be fine-tuned to produce fixed-size, semantically meaningful sequence embeddings optimized for sentence-level similarity tasks. The Transformer-based SEs typically use either the `[CLS]` token embedding or the average of the output token embeddings as the sequence representation. Subsequent methods, such as SimCSE (Gao et al., 2021), showed that more robust sequence embeddings can be learned using contrastive learning objectives, where repeated re-encoding of the same input with different dropout masks provides a self-supervised training signal. Most recently, LLMs have become the foundation for state-of-the-art sequence encoders. LLMs-based SEs leverage supervised contrastive learning combined with instruction fine-tuning, sequence classification, retrieval, or synthetically generated data, achieving leading performance on tasks such as text classification, semantic retrieval, and semantical textual similarity (Muennighoff et al., 2023; BehnamGhader et al., 2024; Wang et al., 2024a).

2.3 Multilingual Representation Learning

2.3.1 Multilingual Pretraining

Multilingual pretraining extends LM training to multiple languages by leveraging large and diverse multilingual corpora, enabling models to capture linguistic nuances and structural properties of different languages, with the goal of inducing a shared multilingual representation space in which semantically equivalent or closely related linguistic units (e.g., words, phrases) across languages exhibit high representational similarity. In this setting, MLM pretraining of Transformer-based LMs is scaled to tens or even hundreds of languages. In what follows, we introduce two prominent encoder-based multilingual pretrained language models (mPLMs): mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020).

Multilingual BERT (mBERT). mBERT is a multilingual extension of BERT, pretrained on Wikipedia text from 104 languages using MLM and NSP. The model employs a shared WordPiece (Wu et al., 2016) vocabulary across all languages and is trained without any explicit cross-lingual supervision or alignment, such as parallel corpora or translation pairs. However, sharing subword units across languages enables the cross-lingual transfer of learned representations (Karthikeyan et al., 2020). During pretraining, batches contain sentences from multiple languages. To mitigate the imbalance between high-resource and low-resource languages, the authors apply temperature-based sampling to up-sample examples from resource-lean languages and down-sample those from resource-rich

languages. As a result, mBERT has been widely adopted for a variety of multilingual NLP tasks (Wu and Dredze, 2019; Kassner et al., 2021).

Unsupervised Cross-Lingual Representation Learning at Scale (XLM-RoBERTa). Concurrently, Conneau and Lample (2019) introduced XLM, a Transformer-based LM trained on both monolingual and parallel corpora using MLM and a newly proposed objective, Translation Language Modeling (TLM). The TLM objective concatenates parallel sentence pairs form a source and a target language, and randomly masks words in both sentences. This forces the model to attend across languages when predicting masked tokens, thereby encouraging stronger cross-lingual alignment at both token and sentence levels. This objective was shown to substantially improve cross-lingual transfer performance.

Building on this work, Conneau et al. (2020) introduced XLM-RoBERTa, which combines the architectural design of RoBERTa (Liu et al., 2019b) with the multilingual training paradigm of XLM (Conneau and Lample, 2019). While following a similar overall approach, XLM-RoBERTa introduces several key modifications. First, it scales self-supervised MLM pretraining to a massive, filtered multilingual Common Crawl corpus (Wenzek et al., 2020) comprising 2.5 TB of text across 100 languages. Second, in contrast to XLM, which leverages parallel data through the TLM objective, XLM-RoBERTa relies solely on MLM with dynamic masking and does not require any parallel corpora. These changes to the training regime result in substantial performance improvements over mBERT and XLM across a wide range of cross-lingual tasks (Conneau et al., 2020).

2.3.2 Multilingual Sentence Encoders

Similar to their monolingual counterparts (cf. § 2.2.2), multilingual SEs can be applied to multi- and cross-lingual sequence-level downstream tasks, such as natural language understanding, semantic similarity, and retrieval. Consequently, a substantial body of research has focused on specializing multilingual encoders to produce sequence-level representations across languages, typically by leveraging large-scale parallel corpora for multilingual alignment (Artetxe and Schwenk, 2019; Yang et al., 2019a; Reimers and Gurevych, 2020a; Feng et al., 2022; Duquenne et al., 2023, *inter alia*).

For example, Artetxe and Schwenk (2019) trained LASER, a sequence-to-sequence LSTM-based LM on a massively multilingual parallel corpus using translation objectives. In their framework, the encoder is shared across languages, while the decoder translates the source-language sentence from a pooled sentence representation produced by the encoder. This design encourages the encoder to learn language-agnostic representations and has been shown to exhibit transfer capabilities even to languages unseen during training. Similarly, Reimers and Gurevych (2020a) trained a multilingual variant of S-BERT by distilling its monolingual counterpart into XLM-RoBERTa using parallel data between English and multiple target languages. Their approach adopts a teacher-student framework, where the student model is trained to mimic the teacher's sentence embeddings while preserving high similarity between translation pairs. This method achieves strong performance on cross-lingual semantic textual similarity and sentence retrieval tasks. In related work, Yang et al. (2020a) trained a universal sentence encoder using a combination of multi-task learning and translation objectives. Feng et al. (2022) proposed LaBSE, a Transformer-based encoder trained in two stages: first, using MLM and TLM objectives, and second, refining the model with contrastive sequence-level objectives. Their training leverages 17B monolingual sentences and 6B parallel sentence pairs, resulting in highly aligned multilingual sentence representations.

More recently, Duquenne et al. (2023) introduced SONAR, a multilingual and multimodal fixed-size sentence embedding model that learns a shared semantic space for text in 200 languages, and extends this space to speech via teacher-student training. SONAR's single encoder-decoder architecture produces language-agnostic sentence embeddings that outperform prior multilingual representations, such as LASER or LaBSE, on both cross-lingual and cross-modal similarity search tasks. Wang et al. (2024b) extended the English E5 embedding models (Wang et al., 2022b) to a multilingual setting by pretraining on 1 billion multilingual text pairs using a contrastive objective,

followed by fine-tuning on labeled datasets to further enhance cross-lingual semantic alignment. Finally, Huang et al. (2025d) proposed a modular multilingual sentence embedding model that separates language specialization from cross-lingual alignment. Their approach first trains language-specific encoders to preserve strong monolingual representations, then introduces lightweight cross-lingual alignment adapters to map non-English embeddings into a shared English-centric space. This modular design achieves more balanced performance across both monolingual and cross-lingual tasks compared to fully shared multilingual encoders.

In Chapter 8 and Chapter 9, we demonstrate that multilingual SEs not only perform strongly in cross-lingual retrieval settings, but can also be easily adapted to domain-specific, sequence-level cross-lingual natural understanding tasks, such as news recommendation. Having discussed the relevant concepts from the field of representation learning in this Chapter, in the next Chapter we introduce the main topic of this thesis: neural news recommendation.

CHAPTER 3

NEURAL NEWS RECOMMENDATION

We shape our tools, and thereafter our tools shape us.

MARSHALL MCLUHAN
«UNDERSTANDING MEDIA: THE EXTENSIONS OF MAN »

In this Chapter, we first provide an overview of personalized neural news recommendation, delving into the core concepts of news and user modeling, ranking, and model training that are relevant for the remainder of the thesis (cf. Section 3.1). In Section 3.2, we outline our evaluation protocol, including evaluation metrics and datasets. Furthermore, in Section 3.3, we discuss open-source frameworks for neural news recommendation and introduce a new library designed to support the development and evaluation of NNRs. Finally, as motivated in Section 1.1, it is equally essential to consider aspects that extend beyond accuracy; therefore, we conclude the Chapter with a discussion of responsible news recommenders and the aspect of multilinguality (cf. Sections 3.4 and 3.5).

3.1 Personalized Neural News Recommendation

Fig. 3.1 illustrates a common framework for personalized news recommendation and its components. The system takes as input a large pool of news articles alongside user data (e.g., click interaction logs). At the core of this pipeline is the recommendation model, which comprises three distinct components – news modeling, user modeling, and a scoring module – and produces a ranked list of articles for each user. The news modeling component forms the backbone of the NNRs; its goal is to encode the content and characteristics of news articles into expressive representations. The user modeling module then leverages these article representations to infer user interests based on interaction history and potential auxiliary signals (e.g., dwell time), ultimately creating a user profile. Subsequently, the scoring component ranks candidate articles in a personalized manner, typically based on their predicted relevance to the user profile. These components (e.g., news encoder, user encoder, scoring module) are usually trained jointly to satisfy specific goals, ranging from accuracy-based objectives (e.g., maximizing clicks) to beyond-accuracy objectives (e.g., increasing diversity). To assess model quality, the generated recommendations are evaluated using accuracy metrics for classification and ranking performance and beyond-accuracy metrics for measuring other dimensions such as diversity and fairness. Crucial to this process is the availability of benchmark datasets and frameworks that support reproducible and fair development and evaluation of personalized news recommenders.

To facilitate the discussion in this Chapter and throughout the remainder of the thesis, we introduce the core notation used for personalized news recommendation. Let $\mathcal{N} = \{n_1, n_2, ..., n_N\}$ denote the set of all news articles and $\mathcal{U} = \{u_1, u_2, ..., u_U\}$ the set of all users. A user $u \in \mathcal{U}$ is associated with an interaction history $\mathcal{H}_u = [n_1^u, n_2^u, ..., n_k^u]$, a sequence of k articles previously clicked or engaged with by the user, optionally accompanied by auxiliary behavioral signals (e.g.,

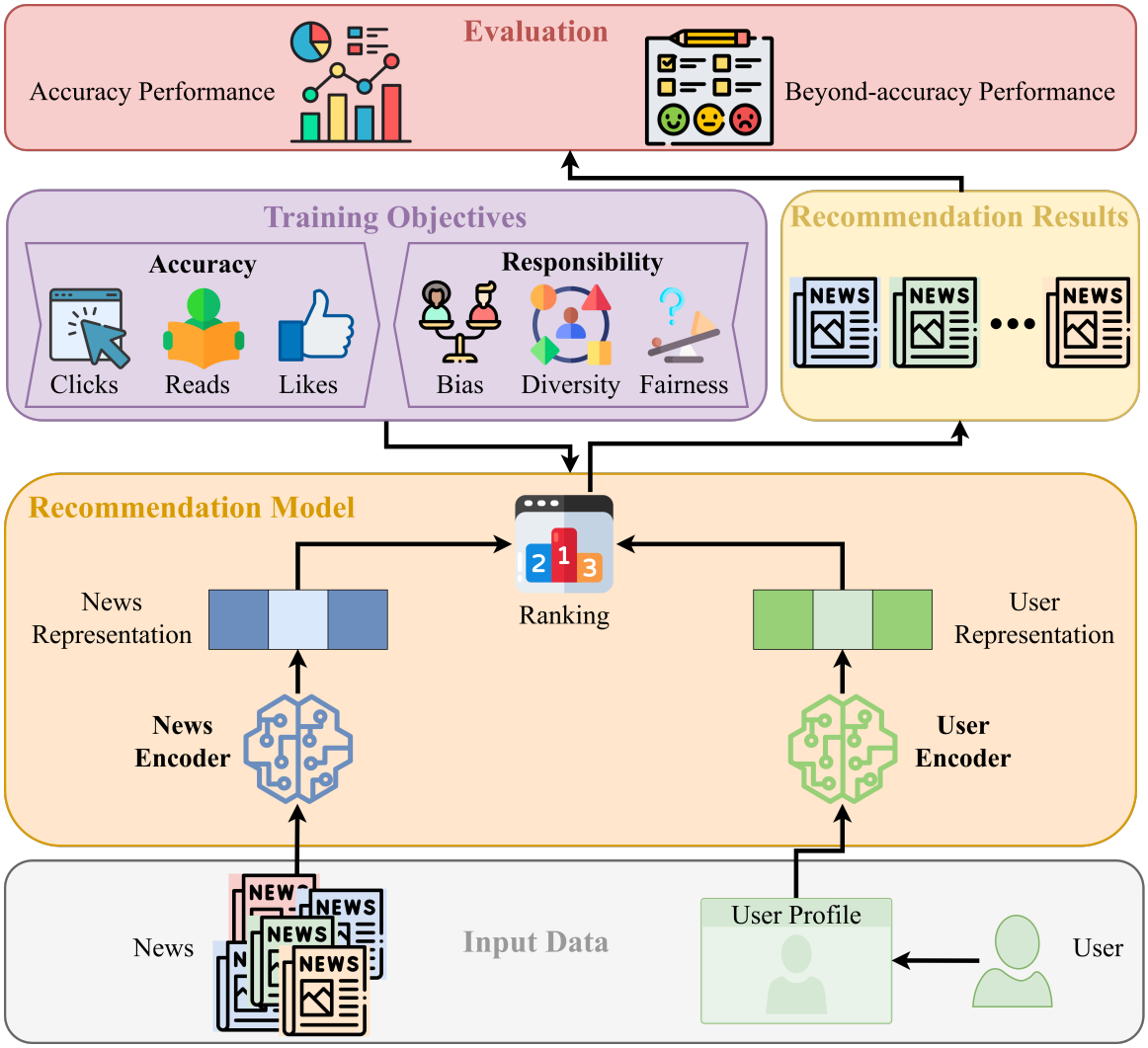


Figure 3.1: Framework of the key components of personalized neural news recommendation models. The figure is adapted from Chapter 4 and Wu et al. (2023).

dwell time), where $n_i^u \in \mathcal{N}$. The news modeling component encodes each news article $n \in \mathcal{N}$ into a vector representation $\mathbf{n}$, while the user modeling component aggregates the representations of the clicked news into a user-level representation $\mathbf{u}$. A scoring function $s(\mathbf{n}^c, \mathbf{u})$ estimates the relevance of a candidate article $n^c \in \mathcal{N}$ for the user u, and the final recommendation list $\mathcal{R}$ consists of candidate articles ranked in descending order of their predicted relevance.

In the following sections, we first discuss relevant techniques for news modeling, user modeling, personalized ranking, and model training. We then provide, in Sections 3.2 and 3.3, an overview of evaluation metrics and benchmarks, as well as a discussion of available frameworks for the development and evaluation of personalized NNRs.

3.1.1 News Modeling

News modeling forms the foundation of any news recommender, aiming to understand and represent the content of news articles from various input features. Following Wu et al. (2023), existing methods can be broadly categorized into feature-based and neural-based approaches. Feature-based methods rely on handcrafted features, such as IDs or manually engineered content features, to encode news characteristics, whereas neural-based methods leverage natural language processing techniques to automatically learn latent article representations directly from raw inputs.

Feature-Based Methods. Feature-based approaches focus on designing or extracting informative attributes from news articles instead of learning news representations end-to-end from scratch (Wu et al., 2023). Early systems, predominantly collaborative-filtering based, represented articles using their IDs (Resnick et al., 1994; Das et al., 2007; Xiao et al., 2015; Ji et al., 2016, *inter alia*). Although computationally simple, ID-based representations are content-agnostic and face significant limitations due to the high churn (i.e. rapid publishing and obsolescence) of news and cold-start issues, as they cannot represent newly published content without interaction history (Wu et al., 2023).

Subsequent works incorporate content-derived features to overcome these limitations. A prominent line of work models textual semantics through bags-of-words or TF-IDF[1] (Sparck Jones, 1972) representations (Gershman et al., 2011), as well as their concept- or synset-aware extensions, such as CF-IDF (Goossen et al., 2011; de Koning et al., 2018; Brocken et al., 2019) or SF-IDF (Capelle et al., 2012; Moerland et al., 2013; Capelle et al., 2015; Hogenboom et al., 2013, *inter alia*). These representations incorporate lexical units, named entities (i.e., real-world objects denoting persons, locations, organizations, etc.), or semantic concepts and synonym sets derived from WordNet (Miller, 1995) to capture news content. Other studies additionally explore additional content features, such as topics (Garcin et al., 2012) or emotional attributes (Parizi and Kazemifard, 2015; Parizi et al., 2016).

Beyond content features, feature-based methods often integrate property and context features (Wu et al., 2023). Property features reflect static, intrinsic article attributes, such as category (Liu et al., 2010b; Li et al., 2011b; Garcin et al., 2012, *inter alia*), publisher (Ilievski and Roy, 2013), or location (Tavakolifard et al., 2013; Kazai et al., 2016; Noh et al., 2014, *inter alia*), with category among the most used attributes to capture news content and user interest. Context features describe temporal or situational dynamic attributes of articles, including popularity and recency (Lee and Park, 2007; Li et al., 2011b,c; Gershman et al., 2011, *inter alia*), novelty (Garcin et al., 2013; Gabrilovich et al., 2004), or publication time (Ilievski and Roy, 2013; Xiao et al., 2015; Chu and Park, 2009; Epure et al., 2017; Fortuna et al., 2010).

While feature-based methods incorporate various types of information, they heavily depend on domain knowledge to design informative features. Consequently, they struggle to capture nuanced linguistic patterns, such word sequences and context within the text, often yielding news representations that are less expressive than those automatically learned through neural-based techniques.

Neural-Based Methods. The advent of deep learning and neural NLP techniques has led to the widespread adoption of neural architectures for news modeling in modern news recommender systems. These approaches learn article representations directly from raw text, alleviating the need for manually designed features. Following the evolution of representation learning and NLP methods (cf. Section 2.1), the majority of neural news modeling approaches can be grouped into three categories: (i) methods using static pretrained word embeddings contextualized via neural networks, (ii) methods leveraging pretrained language models, and (iii) recent approaches based on large language models.

Contextualized Word Embeddings. Early neural approaches relied on static word embeddings, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) to encode individual words in a news article (Khattar et al., 2018; Wu et al., 2019b,a, 2020b, *inter alia*). These word embeddings are then contextualized typically using convolutional neural networks (CNN, Kim, 2014) or self-attention mechanisms (Vaswani et al., 2017). Let the text (e.g, title) of a news article $n \in \mathcal{N}$ be a sequence of words $[w_1, w_2, ..., w_M]$ of length M. This sequence of words is then converted into a sequence of word vectors $[\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M]$ using an embedding matrix $\mathbf{W} \in \mathbb{R}^{V \times D}$, where V is the vocabulary size and D the word embedding dimension.

A prominent line of work uses CNNs to model local contextual patterns around each word. The

[1] Under the Term Frequency–Inverse Document Frequency (TF-IDF) approach (Sparck Jones, 1972), documents are represented as weighted bags of terms, where each term weight reflects its frequency within the document (term frequency) scaled by its inverse frequency across the corpus (inverse document frequency).

contextual representation of the i^{th} word is computed as follows:

$$\mathbf{c}_i = \text{ReLU}(\mathbf{F} \times \mathbf{w}_{(i-K):(i+K)} + \mathbf{b}) \tag{3.1}$$

where $\mathbf{w}_{(i-K):(i+K)}$ denotes the concatenation of the word embeddings within a window size $2K+1$ from position $(i-K)$ to position $(i+K)$, $\mathbf{F} \in \mathbb{R}^{N_f \times (2K+1)D}$ and $\mathbf{b} \in \mathbb{R}^{N_f}$ are the kernel and bias parameters of the CNN filters, N_f is the number of CNN filters, and ReLU (Glorot et al., 2011) is the non-linear activation function. The output of the CNN layer comprises a sequence of contextual word embeddings $[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_M]$. This CNN-based approach is leveraged by numerous widely used NNRs (Wu et al., 2019b,a; An et al., 2019; Zhang et al., 2021c, *inter alia*).

Alternatively, multi-head self-attention (Vaswani et al., 2017) can be used to capture interactions between multiple words and longer-range dependencies among words (Wu et al., 2019e, 2020d, 2022d; Tran et al., 2021). The embedding of the i^{th} word learned by the k^{th} attention head is computed as:

$$\alpha_{i,j}^k = \frac{\exp(\mathbf{w}_i^T \mathbf{Q}_k^w \mathbf{w}_j)}{\sum_{m=1}^{M} \exp(\mathbf{w}_i^T \mathbf{Q}_k^w \mathbf{n}_m)} \tag{3.2}$$

$$\mathbf{h}_{i,k}^w = \mathbf{V}_k^w (\sum_{j=1}^{M} \alpha_{i,j}^k \mathbf{w}_j) \tag{3.3}$$

where $\alpha_{i,j}^k$ denotes the relative importance of the interaction between the i^{th} and the j^{th} words, and $\mathbf{Q}_k^w$ and $\mathbf{V}_k^w$ are the projection parameters in the k^{th} self-attention head. The final multi-head representation $\mathbf{c}_i$ of the i^{th} word is the concatenation of the representations generated by the h separate self-attention heads, i.e., $\mathbf{c}_i = [\mathbf{h}_{i,1}^w, \mathbf{h}_{i,2}^w, ..., \mathbf{h}_{i,h}^w]$.

As individual words contribute unequally to the semantics of a news article, most methods apply additive attention (Bahdanau et al., 2015) to determine the importance of each word based on their contextual embeddings computed previously (Wu et al., 2019e,a; An et al., 2019; Qi et al., 2020; Wu et al., 2020d, *inter alia*). Concretely, the attention weight of the i^{th} word in the news is computed as:

$$\begin{aligned} \alpha_i^w &= \mathbf{q}_w^T \tanh(\mathbf{V}_w \times \mathbf{c}_i + \mathbf{v}_w) \\ \alpha_i^w &= \frac{\exp(\alpha_i^w)}{\sum_{j=1}^{M} \exp(\alpha_j^w)}, \end{aligned} \tag{3.4}$$

where $\mathbf{V}_w$, $\mathbf{v}_w$ are the projection parameters of the attention network, and $\mathbf{q}_w$ is the query vector. The final news embedding n is the weighted sum of its contextual word embedding:

$$\mathbf{n} = \sum_{i=1}^{M} \alpha_i^w \mathbf{c}_i \tag{3.5}$$

Although widely used, these methods rely on shallow, non-contextual word embeddings, which limits their ability to capture deeper semantic nuances in the news text.

Pretrained Language Models. More recent NNRs leverage PLMs, such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019b), to obtain richer and context-aware news representations (Wu et al., 2021b; Yi et al., 2021b; Wu et al., 2021e; Li et al., 2022; Zhang et al., 2024a, *inter alia*). For instance, the PLM-NR framework proposed by Wu et al. (2021b) systematically replaces the original news encoders in established models – i.e., NPA (Wu et al., 2019b), NAML (Wu et al., 2019a), NRMS (Wu et al., 2019e), LSTUR (An et al., 2019)) – with a range of monolingual (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), UniLM (Bao et al., 2020)) or multilingual (e.g., MUSE (Lee and Chen, 2017), Unicoder (Huang et al., 2019), InfoXLM (Chi et al., 2021)) PLMs to improve semantic modeling and enable multilingual news recommendation. Under this paradigm, token-level

representations output by the Transformer encoder are aggregated using attention pooling. Similar methods are used in Zhang et al. (2021b). Other approaches directly use the `[CLS]` token embedding as the news representation $\mathbf{n}$ (Jia et al., 2021; Li et al., 2022; Xiao et al., 2022).

While the use of PLMs substantially improves the semantic quality of news representations, it also introduces considerable computational overhead. In large-scale recommendation pipelines, where thousands of candidate articles must be frequently encoded, this results in significant latency and resource bottlenecks (Xiao et al., 2022). Furthermore, because most PLMs are pretrained on massive, general-purpose text corpora, their learned linguistic patterns might not align well with the stylistic and structural characteristics of news articles, which can lead to suboptimal modeling of news content (Yu et al., 2022).

Large Language Models. Building on the use of contextualized word embeddings and PLMs, recent NNRs increasingly incorporate LLMs to model news content directly from natural language. Owing to their strong language understanding and generative capabilities, LLMs have been adopted either as powerful content encoders or as tools for content enrichment and data augmentation.

The first line of research leverages LLMs to obtain news representations (Zhang and Wang, 2023; Li et al., 2023d,c; Qin et al., 2025), either by prompting them in a zero-shot manner or by fine-tuning them on news-specific data. For example, Zhang and Wang (2023) exploit the reasoning capabilities of LLMs through prompt engineering: they provide the model with both the candidate news article and a user's previously clicked articles, and ask it to infer whether the user would find the candidate article relevant or interesting. Similarly, Li et al. (2023d) design prompt templates that feed structured article information (e.g., title, category) into the LLM to derive content representations.

The second category uses LLMs to (i) generate richer or complementary features that enhance article representations produced by traditional neural encoders (Gao et al., 2024; Runfeng et al., 2023; Kieu et al., 2025a, *inter alia*) or (ii) augment the available training data (Liu et al., 2024b). For instance, Gao et al. (2024) generate theme-level representations of news by prompting an LLM to summarize the central theme of each article; this representation is then fused with semantic features using an attention network. Yada and Yamana (2024) incorporate LLM-generated category descriptions produced by GPT-4 (OpenAI et al., 2024) to enrich article representations derived from titles. Kieu et al. (2025a) propose a hierarchical prompting strategy: first prompting the LLM to generate engaging and contextually appropriate titles, then extracting key entities, and finally combining these elements to produce refined and more compelling article titles.

A comprehensive investigation of LLM usage in both encoding and data generation is presented by Liu et al. (2024b). They use open-source LLMs as content encoders to obtain continuous latent representations based on natural language templates that take as input multiple article attributes (e.g., title, abstract, category). They further fine-tune only the upper layers of the LLM to adapt it to news recommendation. Concurrently, they introduce a generative recommendation framework that employs closed-source LLMs to augment the training data. In this setup, they also prompt the LLM to summarize article content from its attributes, and these enriched summaries are then used as input features for the downstream content encoder.

While LLM-based approaches yield richer news representations, they remain difficult to deploy in large-scale, real-time news recommendation systems due to their inference latency, substantial memory requirements, and strict context-length limitations. Moreover, LLMs often suffer from domain shift and outdated internal knowledge when applied to fast-evolving, time-sensitive content such as news. Finally, their generative nature introduces risks such as hallucination, where events or details are fabricated or misrepresented (Rawte et al., 2023; Ji et al., 2023; Huang et al., 2025b).

Types of News Features. Most NNRs encode news articles primarily using their titles and, in some cases, short abstracts, as these components are considered most influential in a user's click decision. However, several works additionally exploit the full article body to obtain richer semantic representations (Okura et al., 2017; Kumar et al., 2017; Khattar et al., 2018; Wu et al., 2019a, 2020d; Zhang et al., 2021a, *inter alia*). Beyond textual semantics, numerous works incorporate additional

types of news information derived from content (e.g., entities, topics, sentiments), context (e.g., popularity, recency), or external knowledge (e.g., knowledge graphs). These complementary signals capture aspects of news articles not readily conveyed through raw text alone and can substantially strengthen news modeling.

Content Information. A wide range of models leverage additional features extracted directly from news content, such as entities, keywords, topic categories, or sentiment. One popular strategy is to treat named entities as textual tokens and integrate them alongside the article's text to enrich news representations (Gao et al., 2018a; Zhu et al., 2019; Chu et al., 2019; Zhang et al., 2019; Qi et al., 2021e,a, *inter alia*). Topic categories constitute another commonly used attribute, providing coarse-grained semantic cues that are useful for inferring user preferences (Wu et al., 2019a; An et al., 2019; Wu et al., 2019c; Sun et al., 2021, *inter alia*). Categories are typically encoded using ID embeddings (Wu et al., 2019a; An et al., 2019) or pretrained word embedding vectors (Park et al., 2017a; Li et al., 2022), and subsequently fused with text-based representations of news via concatenation (An et al., 2019; Tran et al., 2021; Han et al., 2021) or attention mechanisms (Wu et al., 2019a, 2021a; Sun et al., 2021). Besides topical features, several methods incorporate additional content-derived cues such as news sentiment (Wu et al., 2020c) or the images associated with news articles (Wu et al., 2021c; Xun et al., 2021). By integrating multiple aspects characterizing article content, these approaches learn more comprehensive and accurate news representations, ultimately improving recommendation performance.

Context Features. In addition to content information, context features constitute a valuable source of signals for news modeling. Such features include article popularity, freshness, and user dwell time (Qi et al., 2021c; Meng et al., 2021; Ji et al., 2021; Cho et al., 2021, *inter alia*). For example, Qi et al. (2021c) incorporate near real-time Click-Through Rate (CTR), a proxy for article popularity, alongside textual content and entities, under the assumption that popular articles are more likely to be clicked, although their popularity decays over time. Similarly, Meng et al. (2021) use near real-time CTR and the time elapsed between publication and click to model the evolution of a news article's lifecycle (Meng et al., 2021). Other studies consider presentation-related features, such as the display position or size of an article preview on a news webpage, which can influence click behavior and introduce position bias (Yi et al., 2021a).

External Knowledge. External knowledge sources, particularly knowledge graphs (KGs), provide another powerful means of enriching news representations. KGs (Hogan et al., 2021; Iana et al., 2024) are heterogeneous directed networks that encode real-world entities as nodes and their relations as edges, covering either general-purpose domains, e.g., Wikidata (Vrandečić and Krötzsch, 2014), Microsoft Satori (Team, 2013), or focusing on a particular field, e.g., NewsGraph (Liu et al., 2019a). Entities extracted from news texts are first linked to a KG, enabling models to incorporate structured, commonsense knowledge.

Knowledge-aware news recommenders (Iana et al., 2024) commonly rely on knowledge graph embedding (KGE) models, such as TransE (Bordes et al., 2013) or TransD (Ji et al., 2015), to obtain latent representations of entities and relations based on their structural and semantic context (Cai et al., 2018). These KGEs can be learned from entire large-scale KGs (Wu et al., 2020e; Sun et al., 2021), news-specific KGs that also incorporate collaborative signals (Liu et al., 2019a; Ge et al., 2020), or local subgraphs constructed from k-hop neighborhoods around the extracted entities (Wang et al., 2018b; Sheu and Li, 2020). Another line of work learns entity embeddings directly using graph neural networks, such as graph attention networks (Qi et al., 2021a; Liu et al., 2020a). Entity-based representations are typically fused with the embeddings of text and other features using knowledge-aware CNN architectures (Wang et al., 2018b; Sun et al., 2021) or through concatenation followed by non-linear transformations (Liu et al., 2019a; Gao et al., 2018a; Tran et al., 2021). By incorporating structural external knowledge, these methods capture semantic relations and commonsense patterns not present in raw text, enabling more expressive and informative news representations.

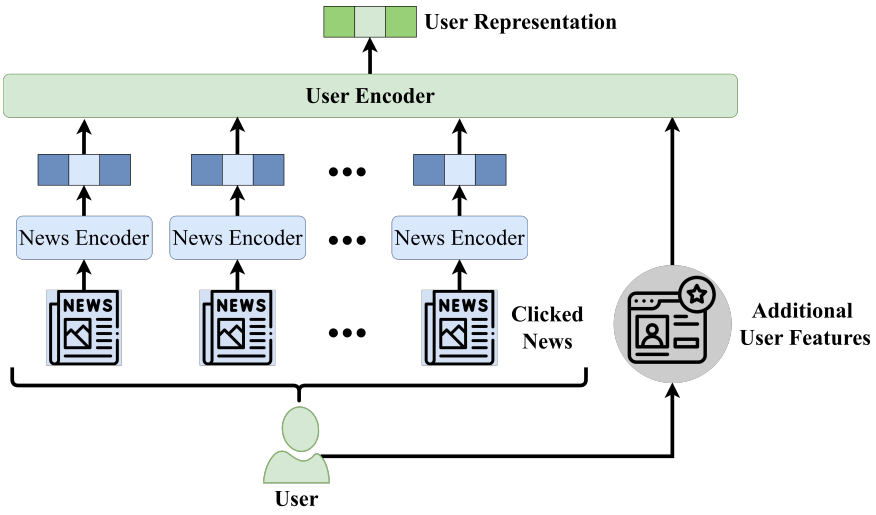


Figure 3.2: An example of user modeling in personalized news recommendation. The figure is adapted from Wu et al. (2023).

Although external data sources such as KGs can enhance representation and address cold-start issues for new articles, their effectiveness depends on accurate entity linking and up-to-date KG coverage. While large, collaboratively maintained resources such as Wikidata or near-real-time KGs derived from Wikipedia (e.g., DBpedia Live) can reflect many emerging entities with limited delay, completeness and consistency across domains and languages is not guaranteed. In particular, local or niche news topics, such as regional events, lesser-known public figures, or community-level organizations, may be underrepresented or incorporated in the KG only after a temporal lag, while entities and concepts expressed in resource-poorer languages often suffer from sparse coverage, inconsistent labeling, or missing cross-lingual links (Conia et al., 2023; Kaffee et al., 2023).

3.1.2 User Modeling

User modeling is the subsequent crucial phase in the personalized news recommendation pipeline. Its primary goal is to infer and encode individual user preferences based on their behaviors. Fig. 3.2 provides a high-level illustrations of the user modeling process: it relies on the news articles a user has interacted with and, optionally, incorporates additional contextual or user-side information to enhance the accuracy of user representations and improve personalization. User preferences can generally be captured through two types of feedback: *explicit* and *implicit* (Knijnenburg et al., 2012). Explicit feedback refers to direct signals provided by users, such as ratings or comments on news articles. In contrast, implicit feedback comprises behavioral cues that can act as proxies for user interests, including clicks, browsing history, reading duration, and scroll depth (Raza and Ding, 2022). As users rarely provide explicit feedback on online news platforms, the vast majority of modern news recommenders rely on implicit feedback for user modeling.

Similar to news modeling techniques, user modeling approaches can be broadly grouped into feature-based and neural-based methods (Wu et al., 2023). Feature-based methods rely on manually engineered features or heuristics to describe users, whereas neural-based methods aim to automatically extract latent patterns from user behaviors to construct rich user profiles.

3.1.2.1 Feature-Based Methods

Feature-based approaches traditionally construct user representations using handcrafted features. Early methods represent users solely by their unique identifiers (Resnick et al., 1994; Das et al., 2007). However, to mitigate the inherent limitations of ID-based collaborative filtering, specifically

data sparsity and the inability to capture user interests in cold-start scenarios, subsequent research incorporated content-based features. For instance, user IDs have been combined with bag-of-words news representations (Lang, 1995) or extracted keywords (Claypool et al., 1999).

Other approaches enrich user representations by incorporating features derived from users' clicked news, including the topics of consumer articles (Li et al., 2011c; Garcin et al., 2012), TF-IDF-based features (Goossen et al., 2011; Capelle et al., 2012), location data (Noh et al., 2014; Asikin and Wörndl, 2014; Robindro et al., 2017, *inter alia*), demographic information (Lee and Park, 2007; Chu and Park, 2009; Yeung and Yang, 2010, *inter alia*), access patterns (Li et al., 2011b), dwell time Yi et al. (2014), or behavior on external platforms, such as social media or search engines (Phelan et al., 2009; Gu et al., 2014; Hsieh et al., 2016; Kazai et al., 2016; Bai et al., 2017, *inter alia*). Another line of work models users via graph-structured data, representing news-user interactions as graphs to capture higher-order relations and better understand user behavior (Li and Li, 2013; Joseph and Jiang, 2019; Iana et al., 2024, *inter alia*).

Despite their historical significance, feature-based methods face notable challenges. Reliance on specific attributes, such as location or demographics, raises privacy concerns and suffers from data unavailability when users withhold personal details. Moreover, these methods depend heavily on domain knowledge to engineer informative features, often resulting in representations that are suboptimal for capturing complex, evolving user preferences. These limitations have motivated the increasing adoption of neural-based methods, which aim to automatically learn expressive user representations directly from raw click behavior. As this thesis focuses on neural news recommender systems, we elaborate on representative neural-based user modeling approaches next.

3.1.2.2 Neural-Based Methods

Neural-based user modeling techniques overcome the reliance on manual feature engineering by automatically identifying and extracting latent patterns from user interaction data (Wu et al., 2023). A wide range of neural architectures have been explored, including attention-based models, CNNs, recurrent networks such as gated recurrent units (GRUs) (Cho et al., 2014) or LSTM networks (Hochreiter and Schmidhuber, 1997), and Transformer-based methods. Despite this architectural variety, these approaches fundamentally differ in how they interpret the sequence of clicked news. Some approaches treat each click as an independent signal of preference, whereas others explicitly capture temporal or semantic dependencies between consecutive interactions. This observation naturally motivates a division between *context-agnostic* and *context-aware* user modeling approaches.

Context-Agnostic User Modeling. A widely adopted class of neural architectures in context-agnostic user modeling relies on attention mechanisms. These methods assume that different clicked articles carry different levels of importance for inferring a user's interests. For example, an article clicked by many users might convey little personal information, whereas a niche sports article may strongly indicate a specific interest. Many representative NNRs (Wu et al., 2019a,c,d; Zhang et al., 2019; Wu et al., 2021b, *inter alia*) employ an additive attention network (Bahdanau et al., 2015) to aggregate clicked news representations. Given the set of clicked news $\mathcal{H}_u = [n_i^u, n_2^u, ..., n_k^u]$, the attention weight α_i^u for news n_i^u is computed as:

$$\begin{aligned} \alpha_i^u &= \mathbf{q}_u^T \tanh(\mathbf{W}_u \times \mathbf{n}_i^u + \mathbf{b}_u) \\ \alpha_i^u &= \frac{\exp(\alpha_i^u)}{\sum_{j=1}^{k} \exp(\alpha_j^u)}, \end{aligned} \tag{3.6}$$

where $\mathbf{q}_u$, $\mathbf{W}_u$, and $\mathbf{b}_u$ are parameters of the attention network, and k denotes the length of the clicked

news history $\mathcal{H}_u$. The user embedding is given by the weighted sum of clicked news representations:

$$\mathbf{u} = \sum_{i=1}^{N} \alpha_i^u \mathbf{n}_i^u \tag{3.7}$$

In these models, the attention query $\mathbf{q}_u$ is typically a global parameter vector shared across users (Wu et al., 2023). Other variants introduce candidate-aware attention mechanisms, where the relevance of each clicked article is conditioned on the candidate news representation, and user embeddings are generated dynamically for each candidate news (Wang et al., 2018b; Qi et al., 2022b; Lee et al., 2020a; Zhang et al., 2021a; Wu et al., 2021c). For example, Wu et al. (2019b) incorporate personalized attention by generating user-specific queries from user ID embeddings. Although such approaches effectively capture informative clicks from user interactions, they ignore dependencies between them, such as temporal order or semantic transitions, that can provide richer insight into user behavior.

Context-Aware User Modeling. To address these limitations, context-aware approaches incorporate structural or sequential information from user click histories.

Sequence-Based Approaches. One line of work models user behaviors as a sequence of interactions. Early examples use CNNs to capture local temporal patterns in user behavior (Zhang et al., 2019; Khattar et al., 2018; Wang et al., 2020). In particular, some methods use 3D CNNs, which treat each article as a 2D matrix obtained by stacking its word embeddings and then order these matrices according to the user's click history (Khattar et al., 2018). This produces a sequence of article representations over which 3D convolutions (Ji et al., 2012) can be applied, allowing filters to simultaneously slide across both the spatial dimensions (capturing local semantic patterns within each article) and the temporal dimension (capturing transitions and dependencies across consecutively clicked articles). This enables the model to encode short-term behavioral context through joint spatiotemporal feature extraction.

More commonly, recurrent neural networks (RNNs) such as GRUs or LSTMs, have been applied to model sequential dependencies (Kumar et al., 2017; Park et al., 2017a; Zhu et al., 2019; An et al., 2019). A representative example is EBNR (Okura et al., 2017), which employs a GRU network over chronologically ordered clicked news. An et al. (2019) further distinguish between short-term preferences, modeled via a GRU, and long-term preferences, captured using user ID embeddings, a technique later adopted also by Tran et al. (2021) and Sun et al. (2021). Formally, given a sequence $\mathcal{H}_u = [n_1^u, n_2^u, ..., n_k^u]$, sorted in ascending order according to the timestamp of clicked news, a GRU updates its hidden states at each time step $t (1 \leq t \leq k)$ as:

$$\begin{aligned}
\mathbf{r}_t &= \sigma(\mathbf{W}_r[\mathbf{h}_{t-1}, \mathbf{n}_t^u]), \\
\mathbf{z}_t &= \sigma(\mathbf{W}_z[\mathbf{h}_{t-1}, \mathbf{n}_t^u]), \\
\tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_{\tilde{h}}[\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{n}_t^u]), \\
\mathbf{h}_t &= \mathbf{z}_t \odot \mathbf{h}_t + (1 - \mathbf{z}_t) \odot \tilde{\mathbf{h}}_t,
\end{aligned} \tag{3.8}$$

where σ is the sigmoid function, $\odot$ the item-wise product, $\mathbf{W}_r$, $\mathbf{W}_z$, and $\mathbf{W}_{\tilde{h}}$ denote the parameter matrices of the GRU network (An et al., 2019). The last hidden state $\mathbf{h}_k$ is used as the short-term user embedding. More complex architectures combine LSTMs with attention mechanisms (Kumar et al., 2017; Zhu et al., 2019), enabling both sequential modeling and importance-weighting of clicked news. While recurrent and convolutional user encoders effectively capture local or sequential patterns, they may struggle to model global structures in consumption behavior (Wu et al., 2023).

Attention-Based Approaches. Inspired by the success of the Transformer architecture (Vaswani et al., 2017), many methods adopt self-attention mechanisms or Transformer-based networks to model global dependencies among clicked news. A representative model is NRMS (Wu et al., 2019e), which captures interactions among clicked news with a news-level multi-head self-attention network,

followed by additive attention that aggregates these representations into the final user embedding. The embedding of the i^{th} news learned by the k^{th} attention head is computed as:

$$\beta_{i,j}^k = \frac{\exp(\mathbf{n}_i^{u^T} \mathbf{Q}_k^n \mathbf{n}_j^u)}{\sum_{m=1}^M \exp(\mathbf{n}_i^{u^T} \mathbf{Q}_k^n \mathbf{n}_m^u)} \tag{3.9}$$

$$\mathbf{h}_{i,k}^n = \mathbf{V}_k^n (\sum_{j=1}^M \beta_{i,j}^k \mathbf{n}_j^u) \tag{3.10}$$

where $\beta_{i,j}^k$ denotes the relative importance between the j^{th} and the i^{th} news in the user history, and $\mathbf{Q}_k^n$ and $\mathbf{V}_k^n$ are the parameters of the k^{th} news self-attention head. The first network outputs the multi-head representation of the i^{th} news as the concatenation of the embeddings generated by the h separate self-attention heads: $\mathbf{h}_i^n = [\mathbf{h}_{i,1}^n; \mathbf{h}_{i,2}^n; ...; \mathbf{h}_{i,h}^n]$ (Wu et al., 2019e). The final user embedding is computed as the weighted sum of the clicked news embeddings:

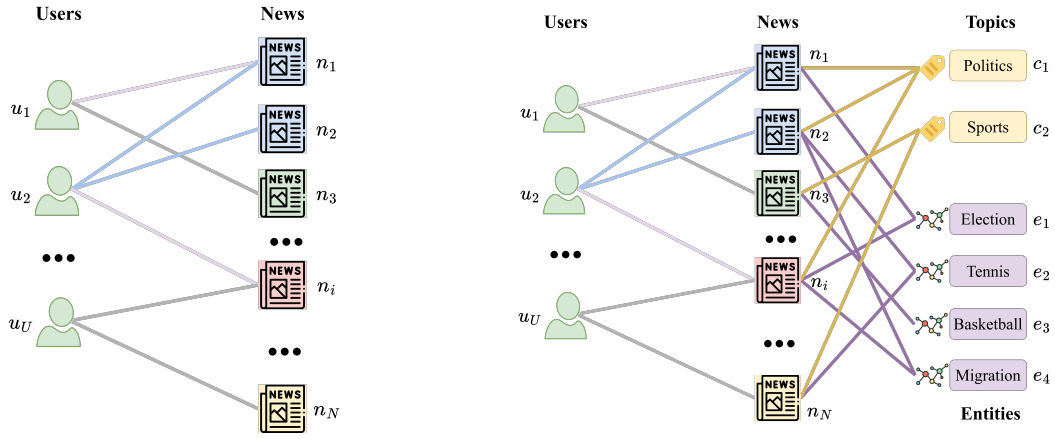
$$\mathbf{u} = \sum_{i=1}^N \alpha_i^u \mathbf{h}_i^n \tag{3.11}$$

where the weight α_i^u is calculated analogously to Eq. (3.6). This user encoder architecture has been widely adopted in several other NNRs (Wu et al., 2021d, 2020c, 2022h; Xun et al., 2021). Other approaches use the `[CLS]` output-token representation of the Transformer as user representations (Zhang et al., 2021b), combine self-attention with candidate-aware attention (Chu et al., 2019) or GRUs (Qi et al., 2020) to enhance user representations, or base more complex user encoders on self-attention networks (Qi et al., 2021a,c). Despite their flexibility, attention-based models may still fail to capture higher-order relational structures among user behaviors (Wu et al., 2023).

Graph-Based Approaches. To capture richer relational patterns, graph-based methods treat user click histories or user-news interactions as graph-structured data, as exemplified in Fig. 3.3 (Wu et al., 2021a; Sheu and Li, 2020; Mao et al., 2021; Zhang et al., 2021c; Tian et al., 2021, *inter alia*). One approach is to model each user's click behaviors as personalized heterogeneous graphs, and to learn user embeddings via graph pooling (Wu et al., 2021a). CAGE (Sheu and Li, 2020) uses a graph convolutional network (Kipf and Welling, 2017) to capture relations between behaviors within a session, combined with a GRU network, whereas Zhang et al. (2021c), Mao et al. (2021), and Tian et al. (2021) model users as entity graphs, and learn latent representations with different types of graph neural networks.

Other approaches additionally leverage collaborative information of users and news (Santosh et al., 2020; Hu et al., 2020a; Ge et al., 2020; Hu et al., 2020b, *inter alia*). For instance, Qian et al. (2019) and Ge et al. (2020) integrate content-based user representations derived from clicked news embeddings with graph-based user representations learned from a user-news graph via graph neural networks. Similar to Zhu et al. (2019) and An et al. (2019), Hu et al. (2020a) learn short-term user representations with RNNs, but instead model long-term user preferences via graph neural networks (Hamilton et al., 2017) applied on a heterogeneous user-news-topic graph.

Graph-based models often show strong performance but typically come with significantly higher computational and memory costs due to the need for iterative message passing and neighborhood aggregation across large user-item graphs. Graph-based architectures materialize multi-hop neighborhoods, intermediate node embeddings, and adjacency structures during training, all of which scale poorly as the graph size grows (Ying et al., 2018; Wu et al., 2022i; Peintner, 2023). Moreover, these methods rely heavily on collaborative signals and therefore struggle to model users who were not observed during training or who have sparse interaction histories. Therefore, this thesis focuses on representative sequential and attention-based user modeling approaches, which offer more favorable scalability properties and more naturally accommodate unseen or sparsely represented users.



(a) A bipartite user-news graph. The figure is adapted from Ge et al. (2020). Edge color indicates different user-news interaction types.

(b) A heterogeneous user-news graph including topic categories and entities. The figure is adapted from Wu et al. (2021a) and Hu et al. (2020a).

Figure 3.3: Examples of user-news graphs used in graph-based user modeling.

LLM-Based Approaches. The advent of LLMs has introduced a new paradigm shift in user modeling, moving from latent vector representations to explicit, semantic, and generative user profiles. Unlike traditional neural methods that learn dense embeddings from click sequences or user IDs, LLM-based approaches leverage the strong semantic understanding and generative capabilities of LLMs to infer user preferences directly from natural language. Existing methods can be broadly categorized into three classes: generative profiling, prompt-based inference, and user simulation.

The first category uses LLMs to summarize user histories into concise, human-readable profiles. Given a user's clicked news, the LLM is prompted to synthesize a thematic description of the user's interests, which is subsequently consumed by downstream ranking modules. For example, Gao et al. (2024) generate high-level semantic profiles from the user interaction histories. Similarly, recent frameworks enhance traditional user representations by generating natural language summaries or by inferring latent topics of interest that might not be directly observable from clicks (Zhang et al., 2024a; Liu et al., 2024b; Zhang et al., 2024b).

A second research direction leverages LLMs' reasoning capabilities through prompt engineering. Here, the user's clicked news, often titles or short snippets, are serialized into a prompt, and the LLM estimates the relevance between a candidate article and the user (Zhang and Wang, 2023; Li et al., 2023b). Prompt4NR (Zhang and Wang, 2023), for instance, formulates news recommendation as a cloze-style prediction task, where the LLM infers a masked action (e.g., related/unrelated, interesting/boring) based on a template that includes a candidate article and the user's history.

The third category employs LLMs as user simulators, addressing challenges such as data sparsity, privacy concerns, and the lack of suitable large-scale datasets (Park, 2025; Xie et al., 2025; Zhang et al., 2025a). These methods use agents that mimic real-world click behavior to generate synthetic interaction data that is then used for training recommender systems. For example, Park (2025) simulates multi-step user sessions by repeatedly sampling candidate news and having an LLM estimate click probabilities based on predefined personas and the user history. Xie et al. (2025) employ an agent-based social simulator to generate group-level social reviews through group deliberation. These synthetic reviews are then combined with the original news text to enhance news and user modeling. Zhang et al. (2025a) similarly adopt an LLM-based agent framework to simulate users, incorporating a stable memory mechanism that captures users' exploration-exploitation dynamic and long-term interest evolution, specifically for the evaluation of news recommender systems.

Overall, LLM-based user modeling offers several advantages: it produces interpretable user profiles, naturally incorporates semantic and contextual information, and generalizes across domains and

modalities. However, these approaches face practical limitations. LLM inference is computationally expensive, potentially limiting scalability in real-time settings. Furthermore, generated profiles may contain hallucinated preferences not grounded in the user history, and transmitting long, sensitive user histories to closed-source LLMs raises privacy concerns.

Types of User Interests. While click behavior is the most widely used implicit signal in many of the NNRs, it can be noisy and not fully reflect user preferences. Consequently, several works incorporate additional information, such as user IDs (Wu et al., 2019b; An et al., 2019) or context features, including time, device, location, or referrer (Gabriel De Souza et al., 2019; Zhang et al., 2019). Other approaches integrate multiple types of user behaviors, including search queries and browsed webpages (Wu et al., 2019d; Shi et al., 2021), dwell time (Wu et al., 2020b), or multi-feedback signals such as click and non-click, share, and dislike (Wu et al., 2022g; Ma et al., 2022).

3.1.3 Personalized Ranking

The third step in the recommendation pipeline is *ranking* the candidate news in a manner that aligns with the user's interests. Wu et al. (2023) categorize ranking techniques into two groups: relevance-based and reinforcement learning-based methods. This thesis focuses on NNRs that adopt a *relevance-based personalized ranking*.

In relevance-based approaches, the system estimates how well each candidate article matches the user's interests by computing a relevance score between their representations, derived from the news and user encoders in earlier stages. Early work measured user-news relevance using cosine similarity between feature vectors (Goossen et al., 2011) or between user and news topic vectors (Garcin et al., 2012). Many neural models adopt the inner product between candidate news and user embeddings as the scoring function (Okura et al., 2017; Wu et al., 2019a; Li et al., 2022, *inter alia*). Other studies have proposed more expressive matching functions, including a feed-forward neural network (Wang et al., 2018b; Hu et al., 2020a), a 3D CNN capturing fine-grained interaction signals between clicked and candidate news (Wang et al., 2020), or hierarchical matching mechanisms that compare candidate articles with user interests at subtopic, coarse-grained, and global levels (Qi et al., 2021e). Möller and Padó (2022) further benchmark several scoring functions against the standard inner product, considering bilinear and non-bilinear formulations, and a two-layer MLP. Across these methods, candidate articles estimated to be more relevant to the user are ranked higher. In the area of responsible news recommendation (see Section 3.4), additional ranking objectives may also be incorporated, such as novelty relative to previously clicked items, or relevance-based lists may be refined using criteria such as recency or popularity (Wu et al., 2023).

The second category of ranking techniques leverages reinforcement learning (RL). Unlike relevance-based ranking, which typically optimizes for immediate objectives such as click prediction, RL-based approaches seek to maximize long-term cumulative reward (Zheng et al., 2018; Song et al., 2021). For example, Li et al. (2010) formulate personalized news recommendation as a contextual bandit problem, while Zheng et al. (2018) employ a Deep Q-Network to estimate policy rewards defined as a weighted combination of click labels and user activeness (derived from their return time after recommendation).

3.1.4 Model Training

The final fundamental component of the recommendation pipeline is the training methodology used to optimize the model, and its modules (e.g., news encoders, user encoders, trainable scoring functions) for accurate recommendation.

Explicit Feedback Optimization. Early news recommender systems relied on explicit feedback (e.g., user ratings) and formulated recommendation as a rating prediction task. These models typically

minimize losses such as the Mean Squared Error (MSE) between predicted and ground-truth ratings:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_{n_i^c} - \hat{y}_{n_i^c})^2 \tag{3.12}$$

where N is the number of training samples, $y_{n_i^c}$ denotes the ground-truth rating, and $\hat{y}_{n_i^c}$ the predicted rating for the i^{th} candidate article (Wu et al., 2023).

Implicit Feedback Optimization However, explicit ratings are scarce in news platforms. Consequently, contemporary NNRs rely predominantly on implicit feedback, particularly click behavior, as a widely available proxy for user interest. As a result, training objectives therefore commonly focus on optimizing the CTR of recommendation results (Wu et al., 2023). This is typically modeled via pointwise, pairwise, or listwise approaches.

Pointwise Binary Classification. Click prediction can be modeled as a binary classification task (e.g., click versus non-click). In this case, each user-item pair is treated independently, with the objective of optimizing the relevance of each candidate directly. The pointwise Binary Cross-Entropy (BCE) loss is widely adopted, framing the task as predicting the probability that a user will click on a specific candidate article (Wang et al., 2018b; Wu et al., 2019b). By minimizing this loss, the model learns to assign high probabilities to clicked candidates and lower probabilities for the ignored ones:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_{n_i^c}\log(p_{n_i^c}) + (1 - y_{n_i^c})\log(1 - p_{n_i^c})] \tag{3.13}$$

where N denotes the number of samples, $y_{n_i^c} \in \{0, 1\}$ is the ground truth label of the i^{th} candidate item (0 for non-click, 1 for click), and $p_{n_i^c}$ is the predicted click probability.

Pairwise Ranking. Alternatively, the problem can be modeled as ranking pairs of items. Bayesian Personalized Ranking (BPR, Rendle et al., 2009) is a popular pairwise ranking loss that encourages the model to rank clicked items higher than non-clicked ones (Qi et al., 2021c, 2022b; Huang et al., 2025c).

$$\mathcal{L}_{BPR} = -\frac{1}{D}\sum_{i=1}^{N}\log\sigma(\hat{y}_{n_i^{c+}} - \hat{y}_{n_i^{c-}}) \tag{3.14}$$

where D represents the training dataset size, σ the sigmoid function, and $\hat{y}_{n_i^{c+}}$ and $\hat{y}_{n_i^{c-}}$ the predicted relevance scores of the i^{th} positive n_i^{c+} and negative n_i^{c-} item, respectively.

Listwise Approaches and Negative Sampling. A substantial number of recent NNRs adopt a listwise approach, maximizing the margin between a positive (clicked) sample and multiple negative (non-clicked) ones simultaneously. A common training objective is the Negative Log-Likelihood (NLL) (Wu et al., 2019b), often implemented as a $(K+1)$ pseudo multi-class classification task over the candidate set using the softmax function:

$$\mathcal{L}_{NLL} = -\sum_{i=1}^{N_p}\log\frac{\exp(\hat{y}_{n_i^{c+}})}{\sum_{j=1}^{K}\exp(\hat{y}_{n_j^{c-}})} \tag{3.15}$$

where N_p denotes the number of positive samples, K the number of associated negative samples, n_j^{c-} is the k^{th} negative sample, and $\hat{y}_{n_i^{c+}}$ and $\hat{y}_{n_j^{c-}}$ indicate the click probability scores of the positive and the negative candidates, respectively. The denominator computes a softmax-normalized click probability. We note that NLL is mathematically equivalent to the standard Cross-Entropy (CE) for $(K+1)$-way classification. Although effective, computing the normalization term in the denominator over the full item set catalog, which often comprises millions of news articles, is computationally

prohibitive for large-scale recommender systems. Therefore, many NNRs resort to restricting the candidate set and adopting alternative objectives.

Noise Contrastive Estimation (NCE, Gutmann and Hyvärinen, 2010) reframes the problem as distinguishing a positive item n_i^{c+} from a set of K negative items $\{n_j^{c-}\}_{j=1}^{K}$:

$$\mathcal{L}_{NCE} = -\sum_{i=1}^{N_p} \log \frac{\exp(\hat{y}_{n_i^{c+}})}{\exp(\hat{y}_{n_i^{c+}}) + \sum_{j=1}^{K} \exp(\hat{y}_{n_{i,j}^{c-}})} \tag{3.16}$$

Since news datasets contain significantly more negatives than positives, training samples are typically constructed by pairing each positive instance with negatives randomly drawn either from the global news pool or negatives from either the full news pool or the specific impression list (Huang et al., 2013; Wu et al., 2022b). The majority of modern NNRs employ this sampling-based loss (Wu et al., 2019a; Zhang et al., 2024b; Ryu et al., 2025; Wu et al., 2025, *inter alia*).

Auxiliary Objectives and Multi-task Learning. Beyond basic click prediction, some models incorporate auxiliary training tasks to enrich news representations or optimize additional objectives. For example, TANR (Wu et al., 2019c) jointly trains on topic prediction to learn topic-aware news representations, while SentiRec (Wu et al., 2020c) integrates sentiment information through a sentiment classification task. Moreover, KRED (Liu et al., 2020a) adopts a multi-task training approach, including item recommendation, item-to-item recommendation, category classification, popularity prediction, and local news detection. Finally, because optimizing solely for clicks may promote undesirable content such as clickbait, some approaches incorporate richer behavioral signals, including dwell time (Yi et al., 2014; Wu et al., 2020d) or finish rate (Wu et al., 2022g), to better capture user satisfaction and long-term engagement.

3.1.5 Selected Baselines

† In this section, we provide an overview of the NNRs that serve as baselines for the majority of experiments conducted in this thesis:

- *DKN* (Wang et al., 2018b) uses a word-entity aligned, knowledge-aware CNN (Kim, 2014) to produce news embeddings from the article titles and their corresponding entities. The entity representations are learned with the TransD (Ji et al., 2015) knowledge graph embedding method. It learns candidate-aware representations of users as weighted sums of clicked news embeddings, where the weights are computed by an attention network that takes as input both the candidate and the clicked news embeddings.

- *NPA* (Wu et al., 2019b) contextualizes pretrained word embeddings with a CNN, followed by a personalized attention module. Its user encoder applies a similar personalized attention module to aggregate the representations of a user's clicked news, using projected user ID embeddings as attention queries.

- *NRMS* (Wu et al., 2019e) learns news representations from pretrained word embeddings through multi-head self-attention (Vaswani et al., 2017) combined with additive attention (Bahdanau et al., 2015) layers, using only article titles as input to the news encoder. It also learns user representations with a two-layer encoder composed of multi-head self-attention and additive attention.

†This Section is adapted from Appendix B of: **Andreea Iana**, Goran Glavaš, and Heiko Paulheim. 2023. News-RecLib: A PyTorchLightning Library for Neural News Recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 296–310, Singapore. Association for Computational Linguistics.

- *NAML* (Wu et al., 2019a) contextualizes pretrained word embeddings using a CNN followed by additive attention (Bahdanau et al., 2015). It additionally incorporates topical information, embedding categories and subcategories through a linear layer. Multiple information sources (e.g., title, abstract, category, subcategory) are combined via additive attention to form the final news embedding. User embeddings are then computed from the clicked news embeddings using another additive attention layer.

- *LSTUR* (An et al., 2019) embeds news in a manner similar to NAML, taking as input the title, abstract, and topical categories. These features representations are concatenated to form the final news embedding. Unlike pervious models, it learns user representations with RNNs: short-term user interest are modeled using a GRU (Cho et al., 2014) over clicked news, and combined with a long-term user embedding that is randomly initialized and fine-tuned. Two variants are proposed to obtain the final user embedding: (i) $\text{LSTUR}_{\text{ini}}$ initializes the GRU hidden state with the long-term user embedding and uses the final hidden state as the user representation; (ii) $\text{LSTUR}_{\text{con}}$ simply concatenates the long- and short-term user embeddings.

- *CAUM* (Qi et al., 2022b) adopts an NRMS-style news encoder, while additionally encoding title entities through attention layers and category information through a dense layer. These features are aggregated to form a unified news representation. Its user encoder combines a candidate-aware self-attention network, which models long-range dependencies among clicked news conditioned on the candidate, with a candidate-aware CNN that captures short-term user interests from adjacent clicks.

- *MINS* (Wang et al., 2022c) embeds textual features (titles and abstracts) using the same architecture as NRMS, and incorporates category information via a linear embedding layer. The final news embedding is produced using additive attention over the different feature embeddings. User representations are learned using an encoder that combines multi-head self-attention, a multi-channel GRU-based recurrent network, and additive attention.

- *CenNewsRec* (Qi et al., 2020) produces contextualized news representations using a CNN combined with multi-head self-attention and additive attention. Its user encoder resembles that of LSTUR, but learns long-term user vectors directly from clicked news via multi-head self-attention followed by attentive pooling, rather than storing explicit user embeddings.

- *MINER* (Li et al., 2022) uses a pretrained BERT model (Devlin et al., 2019) as its news encoder, encoding news titles using the `[CLS]` output-token representation. It also incorporates categorical information using fixed pretrained GloVe embeddings (Pennington et al., 2014) within a category-aware attention weighting module. Its user encoder learns multiple user-interest vectors through a poly attention mechanism that extracts K interests vectors via additive attention. A disagreement regularization encourages interest-vector diversity by minimizing the average pairwise cosine similarity. The matching score between a candidate news item and each interest vector is computed via inner product. Three aggregation strategies are explored: (i) MINER-max, which takes the maximum individual score; (ii) MINER-mean, which averages them; and (iii) MINER-weighted, which uses a target-aware attention network (Wang et al., 2018b) to compute a weighted sum based on the relevance between the candidate and the interest vector.

- *TANR* (Wu et al., 2019c) injects topical-category information by jointly optimizing the NNR for both personalized content recommendation and topic classification. It adopts the same user encoder and news encoder architecture as NAML (Wu et al., 2019a), but category labels are used only for auxiliary prediction, not as embedded inputs to the news encoder, which receives only article titles.

- *SentiRec* (Wu et al., 2020c) uses an architecture similar to NRMS, relying solely on article titles for the news encoder. It injects sentiment information by jointly optimizing for personalized content recommendation and sentiment prediction, and additionally regularizes the model to encourage sentiment diversity.

- *SentiDebias* (Wu et al., 2022h) is a sentiment-debiasing framework built on the NRMS architecture. It employs adversarial learning to reduce sentiment bias originating from user-behavior data and to produce sentiment-agnostic and diverse recommendations.

With the exception of DKN, which uses a feed-forward neural network to compute click probabilities, the remaining NNRs use the dot product between the news and user embeddings to compute relevance or matching scores. Regarding optimization, DKN minimizes the BCE loss, NPA uses the NLL loss, and CAUM adopts BPR. The majority of baselines, namely NAML, NRMS, LSTUR, MINS, and CenNewsRec, are trained using the NCE objective. The multi-objective NNRs augment the recommendation objective (typically NCE) with auxiliary losses via weighted summation. Concretely, TANR adds a CE loss for topic classification. SentiRec integrates two additional terms: the mean absolute error for sentiment classification and a sentiment diversity loss. Similarly, MINER incorporates a disagreement regularization term to minimize the pairwise cosine similarity across interest vectors. Finally, SentiDebias combines NCE with a CE loss for sentiment classification and an orthogonal regularization term, which encourages the decomposition of user interests into sentiment-aware and sentiment-invariant components.

Most earlier models (e.g., NPA, NRMS, NAML, LSTUR) were designed around pretrained word embeddings such as GloVe (Pennington et al., 2014), whereas more recent models (e.g., MINER) already leverage pretrained language models, e.g. BERT (Devlin et al., 2019). As discussed in § 3.1.1, the modern paradigm in neural news recommendation is to replace the original news encoder with PLM-based (or mPLMs-based) encoders to obtain richer semantic representations of news. Accordingly, in Chapters 5 and 7 to 9, following Wu et al. (2021b), we also replace the original news encoders of all recommenders under evaluation with different PLMs or mPLMs.

3.2 Evaluation Protocol

In this thesis, we focus on *impression-based evaluation* of NNRs, in which the gold standard consists of user-candidate pairs: each evaluation instance contains a user and a corresponding list of candidate articles, and each candidate is annotated with a positive or negative label indicating whether it was clicked. Formally, the task input is a set of users $\mathcal{U}$ and a list of candidate news articles $\mathcal{N}^C$. Each user $u \in \mathcal{U}$ is associated with a click history $\mathcal{H}_u = [n_1^u, n_2^u, ..., n_k^u]$, while each article n_i is represented by features such as its title, abstract, or category, as detailed in § 3.1.1. The goal is to generate a ranked list of recommendations $\mathcal{R} = [(n_1^c, s_1), (n_2^c, s_2), ..., (n_M^c, s_M)]$, where each candidate article $n_i^c \in \mathcal{N}^C$ is assigned a relevance score s_i. The recommendation list $\mathcal{R}$ is sorted in descending order of relevance $(s_1 > s_2 > ... > s_M)$, so that articles predicted to be most relevant appear first.

Under impression-based evaluation, each *impression* corresponds to one such recommendation instance: a user, a set of exposed candidate articles – typically including one or more positive (clicked) and several negative (not clicked) items – and ground-truth click labels indicating which articles were clicked and which were not. The performance of a NNR is assessed by how accurately it predicts the candidates' click likelihood, and how effectively it ranks the clicked items above non-clicked ones.

3.2.1 Evaluation Metrics

Accuracy-Based Performance. Throughout this thesis (see Chapters 4 and 9), we primarily use the Area Under the Curve (AUC) as a classification metric to evaluate how well the recommender distinguishes between positive instances (clicked articles) and negative instances (non-clicked articles).

More precisely, AUC corresponds to the probability that a randomly chosen positive item receives a higher relevance score than a randomly chosen negative item:

$$AUC = \frac{\sum_{n_i^{c+} \in N_p} \sum_{n_j^{c-} \in N_n} I(s(n_i^{c+}) > s(n_j^{c-}))}{|N_p| \times |N_n|} \tag{3.17}$$

where N_p and N_n denote the sets of positive and negative samples, and $s(n_i^{c+})$ and $s(n_j^{c-})$ the predicted relevance scores of the positive $n_i^{c+} \in N_p$ and the negative $n_j^{c-} \in N_n$ instances, respectively. The indicator function $I(\cdot)$ evaluates whether a positive sample is scored higher than a negative one:

$$I(s(n_i^{c+}) > s(n_j^{c-})) = \begin{cases} 1, \text{ if } s(n_i^{c+}) > s(n_j^{c-}) \\ 0, \text{ otherwise} \end{cases} \tag{3.18}$$

While AUC assess the discriminative ability, we complement it with ranking-based metrics that evaluate how well the model orders candidate articles within the recommendation list. We use the Mean Reciprocal Rank (MRR) to measure how effectively the system ranks the first relevant (clicked) item in each recommendation list (Voorhees et al., 1999). Let $\mathcal{R}$ denote the recommendation list, and let $rank(n_i^c)$ be the position of the first clicked candidate in $\mathcal{R}$. The MRR is defined as:

$$MRR = \frac{1}{|\mathcal{R}|} \sum_{i=1}^{|\mathcal{R}|} \frac{1}{rank(n_i^c)} \tag{3.19}$$

The reciprocal rank $\frac{1}{rank(n_i^c)}$ penalizes lower positions and yields a value of 1 when the clicked item is ranked first. MRR therefore rewards models that place at least one clicked item near the top of the recommendation list, but it does not account for multiple clicked items.

Therefore, we further employ the Normalized Discounted Cumulative Gain (NDCG), which measures the quality of the ranking by rewarding relevant items appearing near the top of the recommendation list while discounting lower-ranked positions (Järvelin and Kekäläinen, 2002). For a recommendation list truncated at position K, the NDCG is defined as:

$$nDCG@K = \frac{\sum_{i=1}^{K} \frac{2^{rel(n_i^c)}-1}{\log_2(i+1)}}{\sum_{i=1}^{K} \frac{1}{\log_2(i+1)}} \tag{3.20}$$

where $rel(n_i^c)$ is a binary relevance label of the i$^{\text{th}}$ candidate item (typically 1 for clicked, 0 for non-clicked). In contrast to MRR, NDCG naturally accommodates cases where multiple recommended candidate articles are relevant.

Finally, in Chapter 6, we evaluate user engagement using the CTR, which quantities how frequently users click on articles recommended by the system. Formally, it is defined as:

$$CTR = \frac{\text{Number of clicked recommendations}}{\text{Number of recommendations delivered}} \tag{3.21}$$

CTR reflects behavioral engagement rather than the system's ranking quality, thereby complementing the model-centric evaluation offered by AUC, MRR, and NDCG.

Beyond-Accuracy Performance. While accuracy-based metrics assess a recommender's ability to predict clicks or rank candidate articles, they do not capture properties of recommendations that are essential for understanding not only of a recommender's performance, but also its broader implications for users, as discussed in Section 1.1. NNRs can therefore also be evaluated along dimensions such as diversity, novelty, fairness, or serendipity. Since this thesis explores bias effects and diversity in news recommendation, we focus on metrics that quantify the diversity of recommended content.

A standard approach to measuring diversity is to compute the intra-list similarity (ILS), or equivalently, the intra-list distance (ILD), which quantify the average pairwise similarity or distance between items in the recommendation list based on their representations (Ziegler et al., 2005; Castells et al., 2021). Although widely used, these metrics present several limitations in the context of news recommendation. First, ILS/ILD strongly depend on the chosen similarity or distance function, which is tied to the underlying item representations (e.g., TF-IDF vectors, topic distributions, neural embeddings). As highlighted by Vrijenhoek et al. (2022), this leaves the interpretation of what constitutes a diverse list of recommendations open to interpretation and leads to a proliferation of implementations that are dataset and model-specific, limiting comparability and interpretability of diversity scores across studies. Second, these metrics implicitly assume that semantic proximity corresponds to lack of diversity, an assumption that does not necessarily hold true for news. Two articles may be semantically similar in the embedding space while differing along diversity-critical dimensions, such as political leaning, journalistic stance, or source credibility. Conversely, semantically distant articles may nonetheless express highly similar viewpoints. As a result, ILS/ILD may fail to capture the aspects of diversity that matter most from an editorial or societal perspective. Last, these metrics are inherently list-centric: they quantify dissimilarity within the recommendation list but do not consider whether the recommended items introduce new or complementary perspectives relative to what the user has already consumed. For diversity-aware recommendation, central to the analyses in this thesis, it is crucial to evaluate whether recommendations broaden or narrow the user's exposure relative to their historical preferences.

More aspect-specific metrics have been proposed to address some of these limitations. For example, Wu et al. (2020c) introduce $Senti_{MRR}$ and $Senti@k$, which evaluate sentiment diversity between recommended articles and the user's click history. These metrics adapt ranking-based evaluation to sentiment attributes: $Senti_{MRR}$ normalizes reciprocal rank using sentiment scores, while $Senti@k$ is derived from the hit ratio. However, both rely on the average sentiment of a user's clicked items. Consequently, they fail to distinguish between users whose histories share the same average sentiment but exhibit meaningfully different sentiment distributions (e.g., consistently clicking neutral content versus alternating between highly positive and highly negative content).

Given these limitations, this thesis adopts a different evaluation strategy. Following the direction of Vrijenhoek et al. (2021, 2022), we define in Chapter 6 aspect-based diversity and personalization metrics grounded in the distribution of a selected news attributes (e.g, sentiment, topic) within the user's click history and the model's recommendation list. This distributional perspective enables an interpretable, attribute-driven analysis of diversity and bias without relying on predefined similarity functions or coarse aggregate statistics.

3.2.2 Benchmark Datasets

Many models in the field of news recommendation were initially developed and evaluated on proprietary datasets, constructed from user interaction logs and news data collected from platforms such as MSN News (Wu et al., 2019b), Google News (Das et al., 2007), Yahoo! News (Okura et al., 2017), Bing News (Lian et al., 2018), and Sina News (Gu et al., 2016). In contrast, only a limited number of publicly available datasets have been created to support research on personalized news recommendation. Below, we provide an overview of the most widely used public benchmarks:

- **Plista** (Kille et al., 2013) comprises a collection of over 70K German articles gathered from 13 news portals, as well as click data for over 14 million users. It includes article titles and bodies, along with user-related information such as location, device type, browser, and operating system (OS). We note that the Plista dataset is no longer publicly available.

- **Globo**[1] (Gabriel De Souza et al., 2019; de Koning et al., 2018) is a Portuguese dataset collected

[1] https://www.kaggle.com/datasets/gspmoreira/news-portal-user-interactions-by-globocom

from `Globo.com`, a major news portal in Brazil. It contains data for over 300K users and 46K news distributed across 1.2 million sessions. Unlike most public benchmarks, Globo does not provide article text; instead it releases the pretrained Word2vec (Mikolov et al., 2013) embeddings for the news.

- **Adressa**[2] (Gulla et al., 2017) is a Norwegian dataset collected over a three-month period from the Adresseavisen news portal. It is provided in two variants: a *light version* spanning one week with over 11K articles and click data for 561K users, and a *full version* with more than 48K articles and interactions from 3 million users across 10 weeks. In addition to metadata (e.g., categories, authors), the articles include non-disambiguated named entities, while user data contains attributes such as event and reading time, session boundaries, location, subscription status, device type, and OS.
- **MIND**[3] (Wu et al., 2020e) was constructed from user click logs of Microsoft News. It contains 1 million users and over 161K news articles in English. The dataset is enriched with named entities linked to Wikidata, providing more informative features compared to earlier benchmarks. Additionally, the dataset includes 100-dimensional embeddings for entities and relations, pretrained on a Wikidata subgraph (Vrandečić and Krötzsch, 2014) using the TransE knowledge graph embedding method (Bordes et al., 2013). Despite its rich content-level features, MIND provides relatively limited user-side information, consisting only of impression logs, click histories, and timestamps.
- **NPR**[4] (Lucas et al., 2023) is an newer, enhanced version of the Globo dataset, comprising nearly 150K news articles and over 1.4 million impressions from 1.6 million users. Its distinctive feature is the inclusion of the full article text and detailed user behavior logs. Compared to MIND, NPR focuses more on "hard news" such as politics, economy, finance, rather than "soft news" like sports, fashion, entertainment (Vrijenhoek, 2023).
- **EB-NeRD**[5] (Kruse et al., 2024) is a recent Danish dataset collected from the Ekstra Bladet news platform, containing 37 million impression logs covering over 125K news and 1.1 million users. It provides titles, abstracts, and full article bodies, along with categories, sentiment labels, and entities. Similar to Adressa, it offers rich user-level features, including read time, location, subscription status, scroll percentage and demographic information.
- **IDEA**[6] (Heitz et al., 2024) consists of news articles with high topical diversity collected from six UK news outlets. User interaction data was gathered through a two-week user study. While smaller in scale and providing full text only via URLs, it is the most user-informative dataset, including read times, explicit ratings, and scroll percentage, as well as user-level attributes such as political preferences, attitudes towards recommendation algorithms, diversity needs, and satisfaction with the news.

Table 3.1 summarizes the key statistics of these benchmarks, including language, year of collection, dataset size, and available features. Until recently, the Adressa and later the MIND datasets were the only large-scale benchmarks providing sufficiently rich information for training and evaluation, which explains why the majority of prior work relied on them. While newer datasets such as NPR or EB-NeRD have begun to attract increasing attention (Liu et al., 2025; Park, 2025; Daffara et al., 2025; Zhou et al., 2025a, *inter alia*), especially for beyond-accuracy research (e.g., normative diversity) (Kruse et al., 2025; Dattawad et al., 2025; Li et al., 2025; Heitz et al., 2025, *inter alia*), they have not yet achieved the same level of adoption as MIND.

[2] https://reclab.idi.ntnu.no/dataset/
[3] https://msnews.github.io/
[4] https://www.kaggle.com/datasets/joelpl/news-portal-recommendations-npr-by-globo
[5] https://recsys.eb.dk/
[6] https://github.com/Informfully/Datasets/tree/main/IDEA

Dataset	Language	Year	# News	# Users	# Clicks	News Features	User Information
Plista	German	2013	70,533	Unknown[7]	1,095,323	article ID, title, body, publisher, URL, image, timestamp	browser, device, location, do not track, ISP, language, OS, time to action, widget
Globo	Portuguese	2017	46,033	322,897	2,988,181	article ID, category, word embeddings, publisher, timestamp	location, session time, referrer, device type, OS
Adressa	Norwegian	2017	48,486	3,083,438	27,233,576	article ID, title, body, categories, entities, URL, timestamp	user ID, event time read time, session boundary, location, subscription status device type, OS
MIND	English	2019	161,013	1,000,000	24,155,470	article ID, title, abstract, body, categories, subcategories, linked entities, URL, pretrained entity &, relation embddings	user ID, event time
NPR	Portuguese	2023	148,099	1,162,802	1,402,576	news ID, title, body, URL, topics, timestamp	user ID, user type, event time, algorithm, number of clicks, read time, scroll percentage
EB-NeRD	Danish	2023	125,541	1,103,602	37,966,985	article ID, title, abstract, body, categories, subcategories, entities, URL, sentiment, topics, timestamp	user ID, event time, read time, session boundary, location, subscription status, device type, age, gender, scroll percentage
IDEA	English	2023	10,954	593	34,890	article ID, title, abstract body & image URL, author, publisher, timestamp, categories, subcategories	user ID, timestamp, read time, rating, scroll percentage, user background information

Table 3.1: Overview of public news recommendation datasets.

In this thesis, we conduct the majority of our evaluations on the Adressa and MIND datasets, unless specified otherwise. Given their central role in our experimental setup, we provide additional details on both datasets below. Table 3.2 summarizes their main statistics based on the dataset splits obtained after our preprocessing.

Adressa. From the Adressa dataset, we use the *light* (or *1-week*) version, which we refer to simply as Adressa throughout this thesis. Following established practices on splitting the Adressa dataset (Hu et al., 2020b; Xu et al., 2023), we use the data of the first five days to construct user histories and the clicks of the sixth day to build the training dataset. We randomly sample 20% of the last day's clicks to create the validation set, and treat the remaining samples of the last day as the test set. During validation and testing, we reconstruct user histories with all the samples of the first six days of data. Since Adressa contains only positive samples (i.e., clicked articles) and lacks explicit data on non-clicked news, we randomly sample 20 news as negatives for each clicked article to build impressions as proposed by Yi et al. (2021b).

MIND. The MIND dataset is released in two variants, MINDlarge and MINDsmall, the latter constructed by randomly sampling 50,000 users and their interaction logs from the full dataset.[8] We conduct experiments on both MINDsmall and MINDlarge throughout this thesis. Since Wu et al. (2020e) do not release test set labels, we use the respective validation portions for testing, and split the respective training sets into temporally disjoint training (first four days of data) and validation portions (the last day).

[7]Note that in the Plista dataset, users are identified only by their session IDs, and no information is provided to distinguish individual users (Kille et al., 2013). The dataset contains 14,897,978 unique session IDs.

[8]https://github.com/msnews/msnews.github.io/blob/master/assets/doc/introduction.md

	Adressa (one week)			MINDsmall			MINDlarge		
Statistic	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
# News	11,207	11,207	11,207	51,282	51,282	42,416	101,527	101,527	72,023
# Users	83,762	28,141	68,894	45,214	19,703	48,593	643,371	279,853	248,973
# Impressions	182,137	36,711	146,382	124,229	29,498	70,938	1,766,559	420,124	365,201
# Categories	18	18	18	17	17	17	18	18	17
# Subcategories	102	102	102	264	264	252	285	285	269
Avg. history length	12.4	15.6	15.6	33.1	33.9	33.3	33.5	34.5	33.6
Avg. #candidates/user	21.0	21.0	21.0	36.5	40.4	37.5	36.7	40.3	37.4

Table 3.2: Statistics of the Adressa (one week), MINDsmall, and MINDlarge datasets.

3.3 Evaluation Frameworks

‡‡ Despite the abundance of model designs, recommender systems research suffers from two major shortcomings: (i) a surprising amount of non-reproducible research (Ferrari Dacrema et al., 2021) and (ii) unfair model comparisons (Ferrari Dacrema et al., 2019; Sun et al., 2020). The former is, on the one hand, due to many NNR implementations not being publicly released (Sertkan and Neidhardt, 2022a). Existing open source repositories, on the other hand, expose a multitude of programming languages, libraries, and implementation differences, hindering reproducibility and extensibility (Said and Bellogín, 2014). Moreover, a lack of transparency in terms of evaluation datasets, experimental setup and hyperparameter settings, as well as the adoption of ad-hoc evaluation protocols, further severely impede direct model comparisons. Many personalized news recommenders have been evaluated on proprietary, not publicly available datasets, such as Bing News (Wang et al., 2018b), MSN News (Wu et al., 2019a,e), and News App (Qi et al., 2022b)). Even the models trained on the more recently introduced open benchmarks like Adressa (Gulla et al., 2017) or MIND (Wu et al., 2020e)) cannot be directly compared due to the lack of standard dataset splits and evaluation protocols (Wu et al., 2021b; Zhang et al., 2021a; Gong and Zhu, 2022; Wang et al., 2022c). Even more concerning, crucial details regarding the setup of the experiments are regularly omitted from the publications or hard-coded without explanation.

It is thus particularly difficult to evaluate the impact of specific components in NNR architecture and training (e.g., news encoder, user modeling, training objectives) on the overall performance of the model (see Chapter 4). Many models simultaneously change multiple components in both the news and the user encoder, while carrying out only partial ablation studies or evaluating against suboptimal baselines (Rendle et al., 2019).

3.3.1 Open-Source Libraries

In the past decade, numerous frameworks for the development and comprehensive evaluation of recommender systems have been proposed to address the problem of reproducibility in the field (Gantner et al., 2011; Ekstrand et al., 2011; Ekstrand, 2020; Guo et al., 2015; Kula, 2017; Da Costa et al., 2018; Salah et al., 2020; Hug, 2020; Sun et al., 2020; Anelli et al., 2021b). News recommendation poses different challenges for practitioners in comparison to recommendation in domains such as movies, music, or e-commerce (Raza and Ding, 2022; Iana et al., 2024), as discussed in § 1.1.1. However, few of the existing and widely used libraries offer support for news recommenders, and especially for the modern neural news recommendation models.

Microsoft Recommenders (Graham et al., 2019; Argyriou et al., 2020) and RecBole (Zhao et al., 2021, 2022) provide implementation for five and three NNRs, respectively, as well as utilities

‡‡This Section is adapted from: **Andreea Iana**, Goran Glavaš, and Heiko Paulheim. 2023. NewsRecLib: A PyTorch-Lightning Library for Neural News Recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 296–310, Singapore. Association for Computational Linguistics.

Feature / Library	Microsoft Recommenders	RecBole	StreamingRec	Leggomenders	NewsRecLib
Model Support					
# NNRs	5	3	0	7	13
PLM support	✗	✗	✗	✓	✓
Beyond Accuracy					
Models	✗	✗	✗	✗	✓
Metrics	✗	✓	✗	✗	✓
Datasets					
News datasets	MIND	MIND	Plista	MIND, Adressa, EB-NeRD, xMIND	MIND, Adressa, xMIND
System Design					
Modularity	Medium	Medium	Low	High	High
Configurability	Medium	High	Low	High	High
Logging / profiling	Basic	Basic	Basic	Basic	Advanced
LLM Support					
Integration / data augmentation	✗	✗	✗	✓	✗

Table 3.3: Comparison of NewsRecLib against existing open-source libraries that support news recommendation.

for the MIND dataset. Nonetheless, other datasets, more recent approaches, and in particular fairness-aware models and beyond-accuracy metrics are not supported. StreamingRec (Jugovac et al., 2018) is a framework for evaluating streaming-based news recommenders, covering a wide range of algorithms, from trivial baselines (e.g., recently published, most popular) or item-to-item based collaborative filtering or session-based nearest neighbor techniques, to association rule methods and content-based approaches. However, it does not support any of the recent neural models. In these libraries, the sub-modules of a specific recommender are not decoupled from the overall model, which impedes experimentation with and analysis of different model components and training strategies. Legommenders (Liu et al., 2025) is a recent library for content-based recommendation that enables joint training of content encoders alongside behavior and interaction modules, and natively leverages LLMs for data augmentation as well as for content operators and encoders. It includes implementations of several neural news recommendation methods and offers built-in access to commonly used datasets, such as Adressa, MIND, and xMIND (introduced in Chapter 8). Nevertheless, similar to the previously discussed frameworks, Legommenders still lacks functionality for beyond-accuracy evaluation.

Taken together, existing toolkits reveal a clear gap: while several libraries partially support news recommenders, none provide a unified, modular, and reproducible framework tailored to the specific needs of modern NNRs (see Table 3.3). To remedy these critical limitations, and to support the experiments conducted in this thesis, we introduce NewsRecLib[1], an open source library for NNRs. In contrast to the aforementioned frameworks, NewsRecLib focuses solely on the state-of-the-art NNRs, providing utilities for the most used benchmark datasets, architectures, training techniques, and evaluation metrics designed for this domain. NewsRecLib unifies disparate implementations of recent NNRs in a single open-source library that is built on top of mature frameworks for deep learning (PyTorch Lightning), evaluation (TorchMetrics), and configuration (Hydra).

NewsRecLib aims to facilitate reproducible research and comprehensive experimental studies, using an end-to-end pipeline powered by a single configuration file that specifies a complete experiment – from dataset selection and pre-processing over model architecture and training to evaluation protocol and metrics. NewsRecLib is built based on the following guiding principles:

(i) Modularity and extensibility. With PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) as its backbone, NewsRecLib is designed in a modular fashion, with core individual components being decoupled from one another. This enables mixing and matching different modules, as well as seamlessly integrating new ones.

(ii) Easy configurability and reproducibility. NewsRecLib is powered by Hydra (Yadan, 2019), in which each experiment is defined through a single configuration file composed from the configurations

[1]The project can be found at: https://github.com/andreeaiana/newsreclib. The library is licensed under a MIT license.

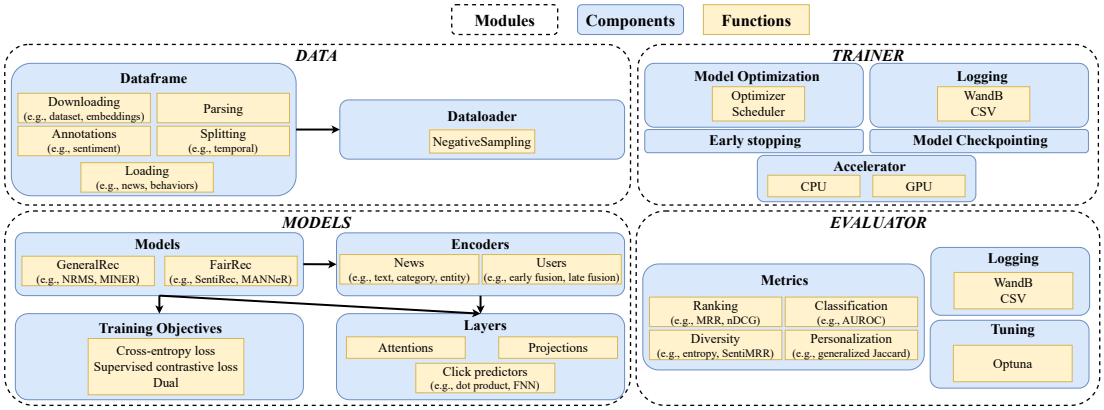


Figure 3.4: Illustration of the NewsRecLib framework.

of specific pipeline components. The configuration of every experiment is automatically stored at the start of the run and as such trivially enables reproducibility.

(iii) Logging and profiling. The library supports multiple standard tools for extensive logging, monitoring, and profiling of experiments with neural models – in terms of losses, evaluation metrics, runtime, memory usage, and model size.

Overall, NewsRecLib is designed to support the development and benchmarking of NNRs as well as the specific analysis of contributions of common components of the neural recommendation pipelines. Next, we discuss the building blocks of NewsRecLib and provide an overview of the readily available models. For a detailed documentation, we refer the reader to the project page.

3.3.2 NewsRecLib

Fig. 3.4 depicts the structure of NewsRecLib, comprising different functional modules: from data modules for downloading and processing datasets to recommendation modules for training and evaluating a particular NNR. The overall pipeline of an experiment is built automatically from the high-level experimental flow provided by the user in the form of a single Hydra configuration file.

Modularization and Extensibility. NewsRecLib is highly modularized: it decouples core components to the largest extent possible. This allows for combinations of different news encoders (e.g., over different input features – text, aspects, entities) with different user modeling techniques, click fusion strategies, and training objectives. NewsRecLib is easily extensible with new features: the user only needs to write a new sub-component class (e.g., category encoder), or, in the case of new datasets or recommenders, to define a new PyTorch Lightning data module or model module, respectively.

Concretely, we decouple the essential building blocks of a NNR, namely the *news encoder*, the *user encoder*, and the *click predictor*. The news encoder is further decomposed into a configurable set of feature encoders (i.e., components that embed different aspects of the news, e.g., title, topical category or named entities). Different model components can be interchanged with corresponding sub-modules of other recommenders, ensuring freedom in choosing each building block of a model independently of the other components (i.e., by mixing the news encoder of "NNR 1" with the user encoder of "NNR 2"), in contrast to practices in existing NNR libraries, in which sub-components are tied to concrete NNR architectures that introduced them. Because of this, NewsRecLib allows for clear-cut and comprehensive analyses of the impact of NNR components on their overall performance.[2] NewsRecLib currently implements feature encoders used in pre-implemented models; users can, however, easily incorporate new ones (e.g., an image encoder) by extending the respective class.

[2]E.g., we leveraged an earlier version of NewsRecLib to analyze the impact of click behavior fusion strategies and training objectives on NNRs' performance (cf. Chapter 4).

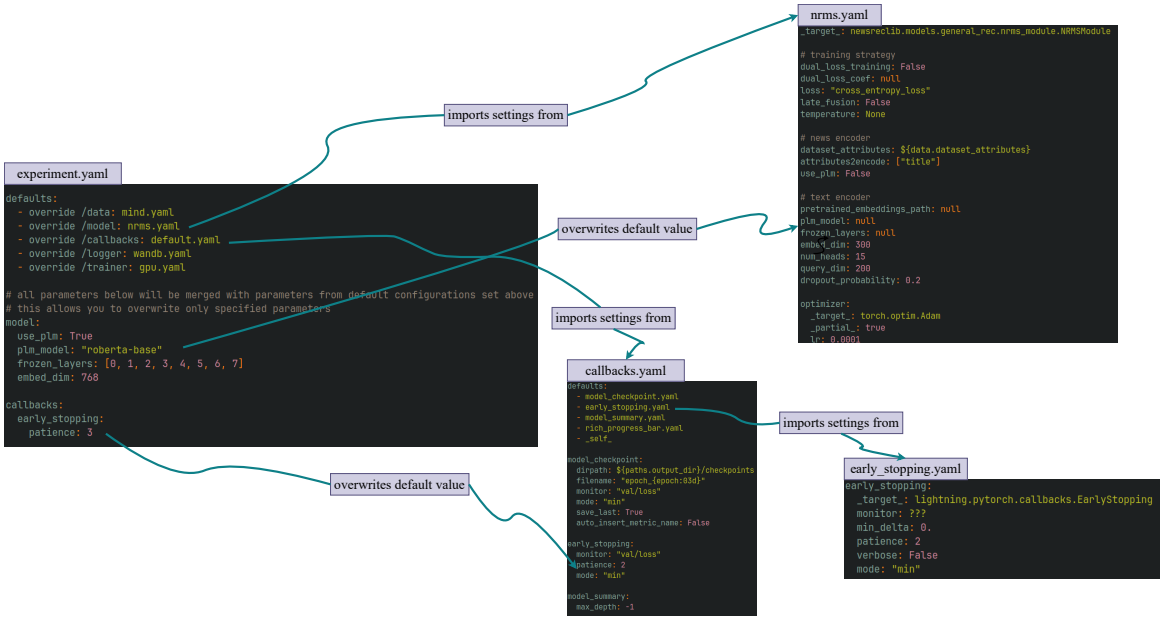


Figure 3.5: A minimal configuration example for training an NRMS (Wu et al., 2019e) model in NewsRecLib. All settings defined in the main and the imported configuration files are merged and persisted into a single configuration object.

Configurability and Reproducibility. Reproducibility strongly relies on the transparency of each step and component in the pipeline, as well as the availability of metadata regarding the factors that influence the model (e.g., hyperparameter values, training objective) and the environment in which it is trained and evaluated (e.g., library versions). Because of this, NewsRecLib leverages the Hydra[3] framework (Yadan, 2019) to decouple the experiment configuration (i.e., a pipeline of modules) from the concrete implementations (i.e., source code) of the modules.

Each concrete module setting is specified and retrieved automatically from a dedicated configuration file which can be accessed by all the pipeline components. A variety of callbacks supported by PyTorch Lightning (e.g., model checkpointing, early stopping, debugging) can be defined, and modified via a corresponding configuration. A single configuration file guides each experiment: the default configurations of the used modules and callbacks are hierarchically inherited and can be overridden. Experiment configurations can also be overwritten directly from the command line, removing the need to store many similar configuration files: this facilitates fast experimentation and minimizes boilerplate code. Experiments can be executed on CPU, GPU, and in a distributed fashion by specifying the type of accelerator supported in PyTorch Lightning. The integration with extensive logging capabilities ensures that any modifications are persistently stored in the experiment directory, together with other log files and model checkpoints.

Fig. 3.5 shows a minimal configuration example for an experiment that trains an instance of the NRMS (Wu et al., 2019e) model. The main configuration file `experiment.yaml` guides the pipeline. It inherits the data and model-specific configurations from `mind.yaml` and `nrms.yaml`, which specify the default configurations of the data module and the NNR model, respectively. `experiment.yaml` further uses the default configurations for the WandB logger, the trainer, and various callbacks. The example also illustrates the interplay between modularization and configurability: we replace the original news encoder of NRMS with a PLM (in this case $\text{RoBERTa}_{\text{Base}}$).

Performance Evaluation and Profiling. With Hydra's pluggable architecture as its backbone, every part of the recommendation pipeline is transparent to the user. NewsRecLib records comprehensive information during training, including number of trainable model parameters and total model size,

[3] https://hydra.cc/

```yaml
defaults:
  - override /hydra/sweeper: optuna

optimized_metric: "val/acc_best"

hydra:
  mode: "MULTIRUN"
  sweeper:
    _target_: hydra_plugins.hydra_optuna_sweeper.optuna_sweeper.OptunaSweeper
    storage: null
    study_name: null
    n_jobs: 1
    direction: maximize
    n_trials: 20
    sampler:
      _target_: optuna.samplers.TPESampler
      seed: 1234
      n_startup_trials: 10
    params:
      data.neg_sampling_ratio: range(1, 10, step=1)
      model.lr: choice(1e-4, 1e-5, 1e-6)
      model.temperature: interval(0.0, 1.0)
```

Listing 1: Example of a hyperparameter optimization process in NewsRecLib. The configuration first runs 10 trials of a search using Bayesian optimization. The hyperparameter search space is defined by indicating the interval, range or choice of values for each desired parameter.

runtimes, training and validation losses. Moreover, it stores important metadata regarding hyperparameter settings, operating system, PyTorch version, environment details, and dependencies between libraries. Any profiler supported by PyTorch can be incorporated by a simple modification of the corresponding configuration file.

NewsRecLib supports widely used loggers like WandB[4] (Biewald, 2020) and Tensorboard[5] (Abadi et al., 2016). Moreover, users can export evaluation metrics for further analysis. Appendix C shows an example of the logging output. We rely on TorchMetrics[6] (Detlefsen et al., 2022) for model evaluation. Users can track numerous metrics ranging from accuracy-based to beyond-accuracy (e.g., diversity) performance. New metrics can be easily added to the pipeline, either by defining the necessary callbacks in the case of metrics already available in TorchMetrics, or by implementing a custom metric as a subclass of the base `Metric` class in TorchMetrics.

Hyperparameter Optimization. NNR performance heavily depends on model hyperparameters, making hyperparameter optimization a crucial ingredient in the empirical evaluation of NNRs. NewsRecLib supports hyperparameter tuning using the Optuna framework (Akiba et al., 2019), which offers a wide range of samplers, such as random search, grid search, and Bayesian optimization (Bergstra et al., 2011; Ozaki et al., 2020).[7] In conjunction with the modularity of NewsRecLib, this allows nearly every component of a news recommender to be treated as a hyperparameter, so that users can optimize the choice of encoders or scoring functions. Listing 1 shows a basic multi-objective hyperparameter search over the number of negative samples, the model's learning rate, and temperature for the supervised contrastive loss.

[4] https://wandb.ai/site
[5] https://www.tensorflow.org/tensorboard
[6] https://torchmetrics.readthedocs.io/en/stable/
[7] https://optuna.readthedocs.io/en/stable/index.html

	Model	News Encoder		Click Behavior Fusion		Training Objective		
		Word emb. + contextualization	PLM	EF	LF	CE	SCL	Dual
GeneralRec	DKN (Wang et al., 2018b)	✓	✗	✓	✓	✓	✓	✓
	NPA (Wu et al., 2019b)	✓	✗	✓	✓	✓	✓	✓
	NRMS (Wu et al., 2019e)	✓	✓	✓	✓	✓	✓	✓
	NAML (Wu et al., 2019a)	✓	✓	✓	✓	✓	✓	✓
	LSTUR (An et al., 2019)	✓	✓	✓	✓	✓	✓	✓
	TANR (Wu et al., 2019c)	✓	✓	✓	✓	✓	✓	✓
	CAUM (Qi et al., 2022b)	✓	✓	✓	✓	✓	✓	✓
	MINS (Wang et al., 2022c)	✓	✓	✓	✓	✓	✓	✓
	CenNewsRec (Qi et al., 2020)	✓	✓	✓	✓	✓	✓	✓
	MINER (Li et al., 2022)	✗	✓	✓	✓	✓	✓	✓
FairRec	SentiRec (Wu et al., 2020c)	✓	✓	✓	✓	✓	✓	✓
	SentiDebias (Wu et al., 2022h)	✓	✓	✓	✓	✓	✗	✗
	MANNeR (cf. Chapter 7)	✗	✓	✓	✓	✗	✓	✗

Table 3.4: List of currently available models in NewsRecLib and supported configurations. For *click behavior fusion* we differentiate between `early fusion` (EF) and `late fusion` (LF). Models can be trained with `cross-entropy loss` (CE), `supervised contrastive loss` (SCL), and a `dual objective` combining both CE and SCL losses as weighted average (Dual). The dashed line separates the general (`GeneralRec`) from the fairness-aware (`FairRec`) recommendation models.

Available Modules. NewsRecLib encompasses three benchmark datasets, 13 news recommendation models, and various evaluation metrics.

Datasets. We provide out-of-the-box utilities for two prominent monolingual news recommendation benchmarks: MIND (Wu et al., 2020e) (with English news) and Adressa (Gulla et al., 2017) (with Norwegian news), as well as for the multilingual dataset xMIND (see Chapter 8). For all datasets, NewsRecLib supports automatic downloading (when available)[8], data parsing, and pre-processing functionalities to create a unified PyTorch Lightning datamodule. Moreover, we include their small and large versions, MINDsmall (xMINDsmall) and MINDlarge (xMINDlarge), and Adressa-1 week and 10 weeks. We additionally automatically annotate the MIND and Adressa datasets with sentiment labels obtained by VADER (Hutto and Gilbert, 2014), a monolingual (English) rule-based algorithm (only for MIND), and a multilingual sentiment classification model of Barbieri et al. (2022a), fine-tuned from XLM-RoBERTa$_{\text{Base}}$ (Conneau et al., 2020).

Recommendation Models. NewsRecLib provides implementations for 10 general-purpose NNRs: (1) *DKN* (Wang et al., 2018b), (2) *NPA* (Wu et al., 2019b), (3) *NAML* (Wu et al., 2019a), (4) *NRMS* (Wu et al., 2019e), (5) *LSTUR* (An et al., 2019), (6) *TANR* (Wu et al., 2019c), (7) CAUM (Qi et al., 2022b), (8) MINS (Wang et al., 2022c), (9) CenNewsRec (Qi et al., 2020), (10) MINER (Li et al., 2022); and 3 fairness-aware recommenders: (11) SentiRec (Wu et al., 2020c), (12) SentiDebias (Wu et al., 2022h), and (13) MANNeR (inroduced in Chapter 7).

To support analysis of model components, for the models that did not use PLMs in their news encoders (but rather contextualized embeddings with convolutional or attention layers), we implement an additional variant with a PLM-based news encoder, as proposed in Wu et al. (2021b). Furthermore, models can be trained either with *early fusion*, i.e., learning a parameterized user encoder to aggregate embeddings of news, or the simpler *late fusion* strategy proposed in Chapter 4, which replaces explicit user encoders with parameter-efficient dot products between candidate and clicked news embeddings.

Table 3.4 provides an overview of the supported configurations for the available models. For each model, users can choose the type of news encoder, click behavior fusion, and training objective. Note that for some models, due to the high interdependencies between news and user encoders, it is

[8]Note that for the Adressa dataset, only a limited version of the dataset is available for download. For the full version containing additional features, users should contact the authors, as detailed in `https://reclab.idi.ntnu.no/dataset/`

not possible to easily replace the original news encoder with a PLM-based one without breaking the framework's modularity. Similarly, some models (e.g., MINER, MANNeR) have been designed from the start with a PLM-based news encoder. In both of these cases, we only provide support for the original news encoder. Due to the design of some model architectures, changing the training objective would modify the functionality of the model (e.g., using different loss functions in the `CR-Module` and `A-Module` of MANNeR cf. Chapter 7). In these cases, we only support one training objective.

Training Objectives. Most NNR models are trained with pointwise classification objectives (Wang et al., 2018b; Wu et al., 2019a,e) with negative sampling (Wu et al., 2019b, 2022b). In Chapter 4, we show that contrastive learning constitutes a viable alternative. At the same time, combining pointwise classification with contrastive objectives has been successfully employed in related tasks (Gunel et al., 2020a). We thus implement three training objectives: cross-entropy loss, supervised contrastive loss (Khosla et al., 2020), and a dual objective that is a weighted average between the two.

Evaluation Metrics. NewsRecLib integrates standard accuracy-based metrics (e.g., AUC, MRR, NDCG), as well as aspect-based diversity and aspect-based personalization defined in Chapter 7. The availability of these beyond-accuracy metrics enables multi-faceted evaluation of NNRs.

3.4 Responsible News Recommendation

Much of the early work in personalized news recommendation has centered on improving model accuracy and maximizing engagement metrics such as CTR. However, as discussed in § 1.1.2, systems that optimize solely for predictive performance and user preference satisfaction can exhibit a range of harmful behaviors, including amplifying misinformation, reinforcing existing viewpoints, or disproportionately favoring certain user or content attributes, that may lead to negative societal or ethical consequences (Wu et al., 2023; Raza and Ding, 2022). Such consequences include the narrowing of exposure to diverse perspectives through filter bubbles and echo chambers (Pariser, 2011; Flaxman et al., 2016; Beam, 2014), disparate treatment of specific user groups (Stinson, 2022; Wang et al., 2023), propagation of historical or structural biases (Chen et al., 2023a), and increased ideological or political polarization (Dandekar et al., 2013; Whittaker et al., 2021). In response, a growing body of research has begun to examine and mitigate the broader algorithmic and societal effects of news recommendation systems. The aim of this field is to design news recommenders that better align with societal values and serve users more responsibly.

Responsible news recommendation encompasses multiple interconnected research directions, including fairness, diversity, debiasing, safety, privacy protection, as well as content moderation aimed at mitigating misinformation and fake news (Wu et al., 2023; Shu et al., 2017; Zhou and Zafarani, 2020). As recommendation models learn from behavioral and content data, both of which can contain historical, societal, or platform-specific biases, they risk inheriting and reproducing these biases in their outputs. This may result in the unfair treatment of certain demographic groups (e.g., systematically under-representing content for minority audiences) or skewed exposure to specific types of news (e.g., over-promotion of sensationalist or partisan content). Consequently, fairness-aware news recommendation has emerged as a key area of study, with the goal of developing algorithms that ensure equitable access to high-quality information across diverse user groups (Burke, 2017; Ekstrand et al., 2022; Wang et al., 2023; Wu et al., 2021d).

A closely related research direction concerns diversity in recommended content. Here, the objective extends beyond satisfying users' immediate preferences to cultivating a more balanced and diverse news diet, thereby mitigating risks associated with filter bubbles and reduced viewpoint diversity (Rao et al., 2013; Vrijenhoek et al., 2021; Bernstein et al., 2021). Approaches in this area aim to encourage recommender systems to present users with a broader range of topics, sources, or perspectives, potentially improving the long-term informational value of the system.

In parallel, the increasing reliance on large-scale user data has raised significant privacy concerns.

Because news recommenders routinely process sensitive behavioral histories, privacy-preserving machine learning methods, such as federated learning (McMahan et al., 2017) or differential privacy (Dwork, 2006), have become central to designing systems that protect users' personal information without sacrificing recommendation quality (Qi et al., 2020, 2021d; Yi et al., 2021b; Liu et al., 2023).

A core focus of this thesis is the study of the algorithmic effects of NNRs and the mitigation of undesirable outcomes through diversification of results. Having already discussed the broader landscape of algorithmic effects in § 1.1.2, we now turn to foundational concepts and relevant work in the closely connected domains of debiasing and diversity.

3.4.1 Debiasing

Recommender systems inevitably inherit various forms of bias from both user behavior data and news content. These biases can affect not only how users interact with news items but also the type of information that is ultimately surfaced to them. In the context of news recommendation, such biases may distort user preference estimation, degrade recommendation quality, or even amplify socially undesirable patterns.

User Behavior Bias. User interaction logs are shaped by numerous factors unrelated to intrinsic user preferences. A well-know example is presentation bias, wherein the visual prominence of news items, such as their position or size on a webpage, substantially influences click behavior. Users tend to click items placed higher in a list or displayed more prominently. Similarly, popularity bias leads users to prefer well-known or trending articles, thereby reducing the visibility of niche or long-tail items. Such behavioral biases introduce noisy signals into training data and may ultimately distort user modeling and degrade recommendation performance (Wu et al., 2023).

To address these challenges, several works propose explicit debiasing strategies. DebiasRec (Yi et al., 2021a), for instance, reduces the influence of position and size bias through a framework comprising three modules: (i) a bias representation model that learns bias representations from features such as article position and display size, (ii) a bias-aware user encoder that extracts unbiased user representations by distilling them from biased histories, and (iii) a click prediction function that decomposes each click score into a bias term and a bias-independent preference term, with only the latter used for ranking. Similarly, DebiasGAN (Wu et al., 2022a) tackles position bias using adversarial learning. The method jointly trains a bias-aware model capturing the effects of article position and a bias-invariant model that estimates position-free click scores using randomized candidate position. Adversarial constraints encourage the user embeddings learned from both models to become indistinguishable, yielding position-robust predictions.

Addressing popularity bias, PP-Rec (Qi et al., 2021c) learns calibrated user interest embeddings by separately modeling article popularity and user preference. The final click score combines a personalized matching score with a popularity score derived from article content, recency, and real-time CTR. More recently, Azevedo et al. (2024) proposed a temporal-counterfactual approach that adjusts training and negative sampling strategies to alleviate popularity bias, improving both accuracy and recommendation diversity.

Content Bias. Biases may also arise from the content of news articles. The style, framing, or tone in which news is written often reflects the perspectives or ideological leanings of publishers, journalists, or affiliated media organizations. When news recommenders are optimized to align with users' historical preferences, they may inadvertently amplify such biases, reinforcing skewed narratives or ideological viewpoints (Liu et al., 2021b; Zhang et al., 2023; Ruan et al., 2023, 2024a).

A growing line of research explores interventions to reduce such content-driven bias. Ruan et al. (2024b) propose an automated rewriting approach to mitigate media bias in news recommendation. Their method uses word-level replacement strategies and LLMs to generate more neutral version of news articles while preserving semantic content. These rewritten articles can then be integrated in the recommendation pipeline at multiple stages, such as into user histories, candidate sets, or training data,

to reduce exposure to biased framing. Wu et al. (2022h) introduce a sentiment-debiasing framework based on decomposed adversarial learning. Their model separates user and news representations into bias-aware and debiased components. During training, both a debiased ranking score (based on bias-agnostic embeddings) and a bias-aware score are used to learn sentiment-related patterns; during inference, only the debiased score is used to rank candidates.

Rather than removing bias, Patankar et al. (2019) estimate a bias score for each article, and use it to present users with alternative, topic-related articles that differ in bias intensity from the selected one, encouraging broader exposure without modifying the original content. Finally, recent research has begun examining cognitive biases introduced when LLMs are used within the news recommendation pipeline. Lyu et al. (2024b) study anchoring (Tversky and Kahneman, 1974), framing (Tversky and Kahneman, 1981), status quo (Samuelson and Zeckhauser, 1988), and group attribution (Hamilton and Gifford, 1976) biases emerging in LLM-based selection and ranking stages. Moreover, they propose several mitigation strategies to counteract the effects of such biases, including synthetic data augmentation, debiasing prompts, or algorithmic adjustments based on human feedback.

Despite these contributions, research specifically targeting critical aspects such as sentiment, stance, frame, or political biases in neural news recommendation remains relatively limited compared to the broader recommender-system literature.

3.4.2 Diversity

Diversity of recommended content constitutes another central consideration in personalized neural news recommendation. Ensuring exposure to a broad range of information is important not only for enhancing user satisfaction, but also for mitigating risks associated with overly narrow and homogeneous news diets, which, in the long run, may undermine public deliberation and democratic discourse (Reuver et al., 2021; Bernstein et al., 2021). Diversity can be incorporated at two main stages of the recommendation pipeline. First, it can be introduced during re-ranking, where the initial list of candidate articles is re-ordered to optimize a diversity-related objective (e.g., maximizing intra-list topical distance or increasing exposure to long-tail content). Second, diversity can be encouraged during model training, by jointly optimizing the recommender for both accuracy and diversity so that recommended items are not only relevant but also reflect multiple facets of the news ecosystem, such as topics, sentiment, frames, political orientations, stances, or perspectives (Raza and Ding, 2022). In this context, diversification may relate to various content attributes, including topical categories, writing style, tone, contextual framing, or ideological positioning (Resnick et al., 2013; Möller et al., 2018; Helberger, 2021), and may consider either a user's historical consumption patterns or the composition of the recommendation list itself (Wu et al., 2023).

Re-Ranking. Early work on beyond-accuracy news recommendation (including non-neural approaches) frequently applied re-ranking techniques to reduce similarity among the final set of recommended items (Ziegler et al., 2005; Rao et al., 2013; Zheng et al., 2018; Tintarev et al., 2018; Gharahighehi and Vens, 2023). For example, Rao et al. (2013) use a news taxonomy to expand a candidate list with related content drawn from encyclopedia websites. Zheng et al. (2018) employ a multi-arm bandit algorithm incorporating exploration-exploitation trade-offs to balance accuracy and diversity, while Li et al. (2011b) re-rank articles within the same topic using popularity-based criteria.

While computationally less demanding than approaches based on joint optimization, diversification through re-ranking comes with several inherent limitations. First, its scope is constrained by the initial candidate set: re-ranking can only rearrange items already retrieved by the recommender and therefore cannot introduce novel content absent from the original list. Second, unlike joint optimization approaches, which simultaneously balance relevance and diversity across the entire item space, re-ranking operates on a pre-filtered set of candidates that are already deemed relevant. As a result, while re-ranking preserves relevance, its ability to meaningfully increase diversity is constrained by the limited variety of the initial recommendations, and substantial diversity gains can

only be achieved when the candidate set itself contains sufficiently diverse items.

Joint Optimization. Recent research increasingly focuses on joint optimization, integrating diversity objectives directly into the training process (Wu et al., 2020c; Shi et al., 2022; Wu et al., 2022c). Raza and Ding (2020) propose combining accuracy and diversity through a dual regularization scheme, using Ridge regression to encourage accurate predictions and Lasso regularization to promote diversity. Wu et al. (2020c) introduce a sentiment-aware NNR that encourages recommendations with sentiment distributions different from those in a user's history. Their method incorporates sentiment cues in news representations via an auxiliary sentiment prediction task and applies a sentiment-diversity regularization technique based on the contrast between the overall sentiment orientation of the user and that of the candidate news. Wu et al. (2022c) present an end-to-end, list-wise diversity-aware model using a permutation Transformer to capture relatedness among candidate items under different permutation orders, thereby learning more varied representations for similar articles. They further employ diversity-aware regularization to ensure that articles with diverse content are placed near the top of the ranked list. Finally, Choi et al. (2022) propose a dual user encoder architecture that distinguishes between generally popular and user-centered preferences. Two complementary regularization strategies, namely similarity regularization and orthogonality regularization, ensure that the former captures broadly shared patterns, while the latter focuses on more discriminative, less popular behaviors, resulting in improved accuracy and more diverse recommendations.

Despite their effectiveness, most of these approaches embed diversification constraints directly into the model architecture or training objectives. As a result, the system learns a single, global balance between accuracy and diversity that is uniformly applied across all users. This universality limits flexibility at inference time and prevents tailoring diversification preferences to individual users (e.g., increasing topical variety for a subset of users or emphasizing stylistic diversity for another) without modifying and retraining the model. Moreover, such methods typically support only single-aspect diversification and struggle to accommodate more complex, multi-aspect requirements (e.g., simultaneously diversifying for topic, sentiment, and framing). Users often value different kinds of diversity – one might seek a broad range of perspectives within a single topic, while another may prefer coverage across many topics but consistent tone or writing style – yet current joint-optimization approaches cannot adapt to these heterogeneous needs without full model retraining.

3.5 Multilinguality

As discussed in § 1.1.2, language is a central dimension of inclusivity in personalized news recommendation. Users who speak multiple languages, or whose languages are underrepresented digitally, face barriers in accessing diverse, relevant, and trustworthy news content. While the motivation for multilingual neural news recommender systems has already been established, this section summarizes the limited existing research on multilingual news recommendation and highlights conceptual foundations from adjacent fields.

Multilinguality in News Recommendation. Despite its importance for inclusive personalized recommendation, research on multilingual news recommendation remains limited. A major contributing factor is that high-quality datasets for news recommendation are overwhelmingly monolingual and primarily cover high-resource languages. Prominent datasets such as Plista (German, Kille et al., 2013), Adressa (Norwegian, Gulla et al., 2017), Globo (Portuguese, de Koning et al., 2018; Gabriel De Souza et al., 2019), MIND (English, Wu et al., 2020e), and EB-NeRD (Danish, Kruse et al., 2024) support only monolingual modeling and evaluation. One notable exception is the multilingual dataset introduced by Wu et al. (2021b), which covers seven languages. However, this dataset is proprietary and restricted to relatively high-resource languages (e.g., French, Italian, Japanese).

Consequently, most early NNRs focused on capturing user preferences and news semantics within a single language. Only a small number of earlier, mostly non-neural, systems explored multilinguality

directly in the context of news recommendation. Among the earliest examples, Yang et al. (2008) propose a cross-lingual framework that recommends, to users browsing news in one language, related articles written in another. Their central idea is to exploit clusters of articles describing the same news event in the source language and then identify corresponding clusters in the target language, thereby providing users with a multilingual perspective over the news event. The system alternates between two learning phases to discover mappings between source- and target-language clusters. In the reverse-mapping phase, classifiers trained on source-language clusters are used to assign target-language clusters (treated as unlabeled data) to one or more source clusters. In the forward-mapping phase, the inferred correspondences enrich the feature representations of the target clusters, and the classifiers are retrained to map from source to target, producing the final cross-lingual mappings. This iterative cross-training procedure yields cluster-level alignments that enable effective cross-lingual, multi-source news consumption. A related but conceptually distinct approach is proposed by Ferdous and Ali (2017), who develop a Bengali-English semantic news recommender grounded in ontology alignment. In their approach, ontologies (Gruber, 1995) – structured representations of concepts and relations extracted from news articles – are constructed independently for each language. The two ontologies are then aligned by translating the source-language concepts into English and resolving translation-induced ambiguities using WordNet (Miller, 1995). Recommendations are subsequently generated by computing a matching factor between the aligned ontologies, which quantifies the semantic similarity between articles across the two languages.

More recent work begins to explore multilinguality within neural news recommendation. Wu et al. (2021b) investigate the use of mPLMs as news encoders to enable multilingual news recommendation on a proprietary seven-language dataset. Guo et al. (2023) address few-shot news recommendation by transferring user-news preference patterns from a many-shot source domain (with abundant interaction data) to a few-shot target domain (where limited interactions lead to cold-start issues). Their approach relies on an unsupervised cross-lingual transfer model to align semantically similar article across the two domains, which may differ in both language and topical distribution. The news and user encoders are shared across domains to facilitate this alignment. To mitigate domain shift, they propose three complementary strategies. First, cross-domain extension, which augments each source-domain article with semantically-similar target-domain articles and machine-translated variants. Second, they employ random masking, which randomly removes during training two elements from each triplet (source article, target-domain similar article, target-language translation), encouraging the news encoder to learn unified representations across languages and domains. Third, the news alignment introduces an auxiliary MSE objective to pull the representations of matched (source versus augmented/translated target) news items closer together, thereby enforcing cross-domain semantic alignment. Together, these mechanisms enable effective cross-domain, and crucially, cross-lingual, news recommendation in few-shot settings.

Overall, while these works take important steps toward multilingual news recommendation, the small number of studies and the absence of suitable multilingual datasets underscores that this research area remains largely unexplored, and emphasizes the need for generalizable methods capable of handling multilinguality in more realistic and diverse settings.

Foundations from Multilingual NLP and Cross-lingual IR. Although only a few works directly address multilingual news recommendation, several strands of research in multilingual NLP, information retrieval (IR), and cross-domain recommendation provide methodological foundations relevant to multilingual and cross-lingual personalized NNRs.

Multilingual Representation Learning. A central requirement for multilingual news recommenders is representing news content from different languages in a shared semantic space. Advances in multilingual representation learning, discussed in Section 2.3, particularly pretrained multilingual language models such as mBERT (Devlin et al., 2019) or XLM-RoBERTa (Conneau et al., 2020), as well as multilingual sentence encoders like m-USE (Yang et al., 2020b), LASER (Artetxe and Schwenk, 2019), or LaBSE (Feng et al., 2022), demonstrate that encoders trained with parallel corpora,

contrastive alignment, or MLM objectives can produce language-agnostic text representations with strong cross-lingual generalization. A key advantage of such multilingual encoders is their built-in cross-lingual transfer capability: semantic knowledge learned from high-resource languages can be directly applied to low-resource ones without explicit supervision. In news recommendation, this enables leveraging abundant English or Chinese data to improve representations for resource-leaner languages, and training user encoders that capture multilingual reading patterns even when interaction data in some languages is sparse.

Within neural news recommendation, multilingual encoders allow content similarity to be computed directly across languages, without translating articles or maintaining separate encoders. This unified latent space reduces modeling and engineering overhead, supports cross-lingual generalization, and enables expressive representation of multilingual user behavior. Early evidence of this approach is provided by Wu et al. (2021b), who show that mPLM-based news encoders serve as strong backbones for multilingual NNRs, highlighting the potential of multilingual representation learning for scalable and generalizable multilingual news recommender systems.

While pretrained multilingual encoders provide strong cross-lingual alignment, they are optimized for general-purpose text. News, characterized by dense named entities, temporally anchored events, and domain-specific vocabulary and style, presents unique challenges. Domain-adaptive pretraining (Beltagy et al., 2019; Gururangan et al., 2020, *inter alia*) has been shown to improve downstream performance for semantic similarity and entity-level tasks. In multilingual news recommendation, domain adaptation can enhance cross-lingual article representations and improve recommendation accuracy when user histories span multiple languages.

Cross-Domain and Transfer Learning in Recommendation. Multilingual news recommendation can be framed as a special case of cross-domain recommendation, where each language constitutes a distinct domain with its own content distribution, interaction patterns, and representational challenges. Early cross-domain recommender systems explored transferring knowledge across heterogeneous domains using shared latent factors, domain-specific encoders, and auxiliary alignment objectives (Fernández-Tobías et al., 2012; Zhu et al., 2021a; Zang et al., 2022). These methods demonstrate how user preferences and item schematics learned in one domain can support recommendations in another. Viewed through this lens, languages can be treated as related but distributionally distinct domains. Transfer-learning principles from cross-domain recommendation provide a useful conceptual foundation for multilingual neural news recommendations: parameter sharing, representation alignment, and auxiliary consistency or alignment losses facilitate the transfer of user preference patterns and semantic knowledge from source to target languages.

Cross-Lingual Information Retrieval. Cross-lingual information retrieval (CLIR) provides another relevant conceptual foundation. It investigates how semantically relevant documents in a target language can be retrieved for a query issued in a source language, using techniques such as dictionary- or MT-based query translation, bilingual vector-space alignment, and parallel-corpus-based semantic mapping (Nie, 2010; Galuščáková et al., 2021). These methods demonstrate that cross-lingual retrieval is feasible when content representations can be compared in a shared semantic space.

Although modern multilingual NNRs do not perform retrieval in the classical CLIR sense, the underlying challenge is similar: matching user histories or preferences expressed in one or multiple languages with candidate articles written in another. CLIR highlights challenges relevant for news recommendation, such as lexical divergence, imbalanced resource availability, and cross-lingual topical shifts. Neural approaches, in particular, leverage bi-encoder and cross-encoder architectures built on mPLMs to project queries and documents into shared semantic spaces for end-to-end ranking (Guo et al., 2020a; Yu and Allan, 2020; Litschko et al., 2022, *inter alia*). This perspective underscores a key principle for multilingual personalized news recommendation: content and user representations can be aligned across languages without explicit translation, enabling cross-lingual matching of relevance and preference signals.

Summary. In summary, multilingual news recommendation builds on a combination of approaches from multilingual representation learning, cross-domain transfer, and cross-lingual information retrieval. Pretrained multilingual encoders provide shared semantic spaces for content across languages, transfer learning principles enable the propagation of user preferences between linguistically distinct domains, and insights from CLIR illustrate how cross-lingual relevance can be modeled without explicit translation, Together, these foundations establish a conceptual and methodological basis for developing neural news recommender systems capable of handling multilingual and cross-lingual interactions, even in the absence of large-scale multilingual datasets.

After having discussed the fields of representation learning and neural news recommendation, in the following Chapters we will dive into each of the challenges outlined in § 1.1.3, presenting and discussing the proposed solutions.

Part II

ANALYSIS OF COMMON COMPONENTS

Chapter 4

User Modeling and Training Objectives

There are many paths to the top of the mountain, but the view is always the same.

Chinese Proverb

* As discussed previously (see Section 3.1), the downstream performance of a neural recommender system is shaped by all its constituents building blocks, ranging from news and user modeling to scoring functions and training objectives. Accordingly, a fundamental challenge for the effective development of NNRs lies in disentangling how design choices in these shared components influence recommendation quality, and in identifying principled ways to improve them to achieve more effective and efficient recommendation (**C1**).

The advent of personalized news recommendation has given rise to increasingly complex recommender architectures. Most NNRs (cf. § 3.1.2) rely on user click behavior and typically introduce dedicated user encoders that aggregate the content of clicked news into user embeddings (early fusion). These models are predominantly trained with standard pointwise classification objectives (cf. § 3.1.4). The existing body of work exhibits two main shortcomings: (1) despite general design homogeneity, direct comparisons between models are hindered by varying evaluation datasets and protocols; (2) it leaves alternative model designs and training objectives vastly unexplored.

In this Chapter, we present a unified framework for news recommendation, allowing for a systematic and fair comparison of news recommenders across several crucial design dimensions: (i) candidate-awareness in user modeling, (ii) click behavior fusion, and (iii) training objectives. Our findings challenge the status quo in neural news recommendation. We show that replacing sizable user encoders with parameter-efficient dot products between candidate and clicked news embeddings (late fusion) often yields substantial performance gains. Moreover, our results render contrastive training a viable alternative to pointwise classification objectives.

4.1 Introduction

In recent years, content-based news recommendation has seen increasingly complex neural recommender architectures that aim to customize suggestions to users' interests (Li and Wang, 2019; Wu et al., 2023). Most neural news recommendation models commonly comprise (i) a dedicated news encoder (NE) and (ii) a user encoder (UE) (see Section 3.1). NEs – instantiated as a convolutional

*This Chapter is adapted from: **Andreea Iana**, Goran Glavas, and Heiko Paulheim. 2023. Simplifying Content-Based Neural News Recommendation: On User Modeling and Training Objectives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2384–2388, New York, NY, USA. Association for Computing Machinery.

network (Wang et al., 2018b; Wu et al., 2019a,c), self-attention network (Wu et al., 2019e, 2020c; Qi et al., 2021c), graph attention network (Qi et al., 2021b), or pretrained Transformer network (Wu et al., 2021b; Yu et al., 2022) – convert input features (e.g., titles, categories, entities) into the news embedding (cf. § 3.1.1). UEs aggregate embeddings of clicked news into a user-level representation by means of sequential (An et al., 2019; Qi et al., 2020; Wang et al., 2022c) or attentive (Wu et al., 2019a,e; Wang et al., 2018b) encoders that contextualize embeddings of clicked news based on patterns in clicking behavior (Okura et al., 2017; An et al., 2019; Wu et al., 2022e, § 3.1.2). We dub this predominant paradigm *early fusion* (EF) because it aggregates representations of clicked news (i.e., builds user representation) before comparison with the candidate.

Most NNRs encode users and candidate news separately, in a *candidate-agnostic* manner (An et al., 2019; Wu et al., 2019a,e). *Candidate-aware* models (Wang et al., 2020; Qi et al., 2021e; Zhang et al., 2021a; Qi et al., 2022b, *inter alia*), in contrast, acknowledge that not all clicked news are equally informative w.r.t. the relevance of the candidate (e.g., a candidate is often representative of only a subset of a user's preferences), and contextualize representations of clicked news with the embedding of the candidate in user-level aggregation with UE. Finally, the candidate's embedding (output of NE) is compared against the user embedding (output of UE): the candidate's recommendation score is computed directly as the dot product of the two embeddings (Wu et al., 2019a) or with a feed-forward scorer (Wang et al., 2018b). NNRs are predominantly trained via standard classification objectives (Wang et al., 2018b; Wu et al., 2019a,e, 2021b) with negative sampling (Huang et al., 2013; Wu et al., 2019b).

The existing body of work has two main shortcomings. First, despite general design homogeneity, direct comparisons between recent NNRs are hindered by lack of transparency and adoption of ad-hoc evaluation protocols (Raza and Ding, 2022; Iana et al., 2024). In particular, a vast majority of personalized news recommenders are evaluated on proprietary datasets, such as MSN News (Wu et al., 2019a,e), Bing News (Wang et al., 2018b), and NewsApp (Qi et al., 2022b). Even the few models evaluated using the publicly available datasets like Adressa (Gulla et al., 2017) or MIND (Wu et al., 2020e) cannot be directly compared due to different dataset splits and evaluation protocols (e.g., model selection strategy) (Zhang et al., 2021a; Gong and Zhu, 2022; Wu et al., 2021b; Wang et al., 2022c). Secondly, simpler and arguably more intuitive design alternatives have largely been left unexplored. On the one hand, the existing work adopts EF as default architecture, proposing increasingly complex user encoding components (An et al., 2019; Qi et al., 2022b), often with little empirical justification for added complexity. On the other hand, only a small fraction of NNRs leverage contrastive learning objectives (Wu et al., 2022b; Yu et al., 2022; Tong et al., 2025), despite such training criteria being proven highly effective in closely related retrieval and recommendation tasks (Li et al., 2021; Wei et al., 2021; Yang et al., 2022; Xie et al., 2022, *inter alia*).

In this Chapter, we remedy the above shortcomings of current NNRs and shed new light on user modeling and training objectives.[1] **1)** Concretely, we introduce a unified framework for neural news recommendation, facilitating systematic and fair comparison of NNRs across three crucial design dimensions: (i) *candidate-awareness in user modeling*, (ii) *click behavior fusion*, and (iii) *training objectives*. **2)** We propose to replace user modeling with complex user encoders (i.e., *early fusion*) with simple pooling of dot-product scores between candidate and clicked news embeddings (i.e., *late fusion*). We show that, despite conceptual simplicity, LF brings substantial performance gains over EF-based NNRs, rendering complex UEs empirically unjustified. **3)** Finally, we demonstrate the benefits of supervised contrastive training as a viable alternative to pointwise classification. Our work fundamentally challenges the status quo of NNRs by introducing simpler and more effective alternatives to the established paradigm based on complex user modeling.

[1] **Disclaimer:** In this Chapter we focus exclusively on NNRs that do not resort to graph-based modeling of relations between users.

4.2 Methodology

Fig. 4.1 depicts our unified evaluation framework for neural news recommendation, focusing on three critical dimensions. Given input data, comprising news and user behaviors, we analyze (i) candidate-agnostic (C-AG) versus candidate-aware (C-AW) user modeling under (ii) two click behavior fusion strategies, namely EF and LF, where each model can be (iii) trained by minimizing either the standard cross-entropy loss (CE) or a supervised contrastive objective (SCL). Next, we describe the models selected for evaluation and formalize the concrete design choices.

4.2.1 User Modeling

Candidate-Agnostic (C-AG) Models. For these models, the UE produces the user embedding from embeddings of clicked news without contextualization against the candidate. We evaluate the following C-AG models, mutually differing in their NE component (i.e., how they embed the clicked news): (1) *NPA* (Wu et al., 2019b) uses a personalized attention module to aggregate the representations of the users' clicked news, with projected embeddings of the users IDs as attention queries; (2) *NAML* (Wu et al., 2019a) uses additive attention (Bahdanau et al., 2015) to encode users' preferences; (3) *NRMS* (Wu et al., 2019e) learns user representations with a two-layer encoder that consists of multi-head self-attention (Vaswani et al., 2017) and additive attention; (4) *LSTUR* (Wu et al., 2020e) learns user representations with recurrent networks: a short-term user embedding is produced from the clicked news with a GRU (Cho et al., 2014), and combined with a long-term embedding, consisting of a randomly initialized and fine-tuned part; the final user embedding is then obtained either (i) as the final hidden state of the short-term GRU, initialized with the long-term embedding ($LSTUR_{ini}$), or (ii) by simply concatenating the short- and long-term user embeddings ($LSTUR_{con}$); (5) *CenNewsRec* (Qi et al., 2020) adopts a similar UE architecture as LSTUR, but learns long-term user vectors from clicked news using a sequence of multi-head self-attention and attentive pooling networks, as opposed to storing an explicit embedding per user; (6) *MINS* (Wang et al., 2022c) encodes users through a combination of multi-head self-attention, multi-channel GRU-based recurrent network, and additive attention.

Candidate-Aware (C-AW) Models. UEs in candidate-agnostic models produce the same user embedding, regardless of the content of the candidate news. In contrast, UEs of candidate-aware models, two of which we include in our empirical analysis, produce user embeddings dependent on the candidate. (7) *DKN* (Wang et al., 2018b) computes candidate-aware representations of users as the weighted sum of their clicked news embeddings, with weights being produced by an attention network that takes as input the embeddings of the candidate and of the clicked news, as produced by the NE. More recently, (8) *CAUM* (Qi et al., 2022b) combines (i) a candidate-aware self-attention network to model long-range dependencies between clicked news, conditioned on the candidate, and (ii) a candidate-aware CNN to capture short-term user interests from adjacent clicks, again conditioned with the candidate's content; the candidate-aware user embedding is finally obtained by attending over the long-range and short-term representations.

News Encoders. The NNRs included in our evaluation primarily use news titles as input, which they typically embed via pretrained word embeddings (Pennington et al., 2014). NAML, LSTUR, MINS, and CAUM additionally leverage category information, with categories embedded with a linear layer. CAUM also encodes title entities and DKN exploits knowledge graph embeddings (Ji et al., 2015). The shallow word and entity embeddings are contextualized either using a combination of multi-head self-attention (in NRMS, MINS, CAUM), or a sequence of CNN (Kim, 2014) and additive attention networks (in NAML, LSTUR). NPA (Wu et al., 2019b) also utilizes a CNN to contextualize word embeddings, followed by a personalized attention module, analogous to the one used in its user encoder, whereas DKN employs a word-entity-aligned knowledge-aware CNN (Wang et al., 2018b). CenNewsRec (Qi et al., 2020) combines the CNN with multi-head self-attention and

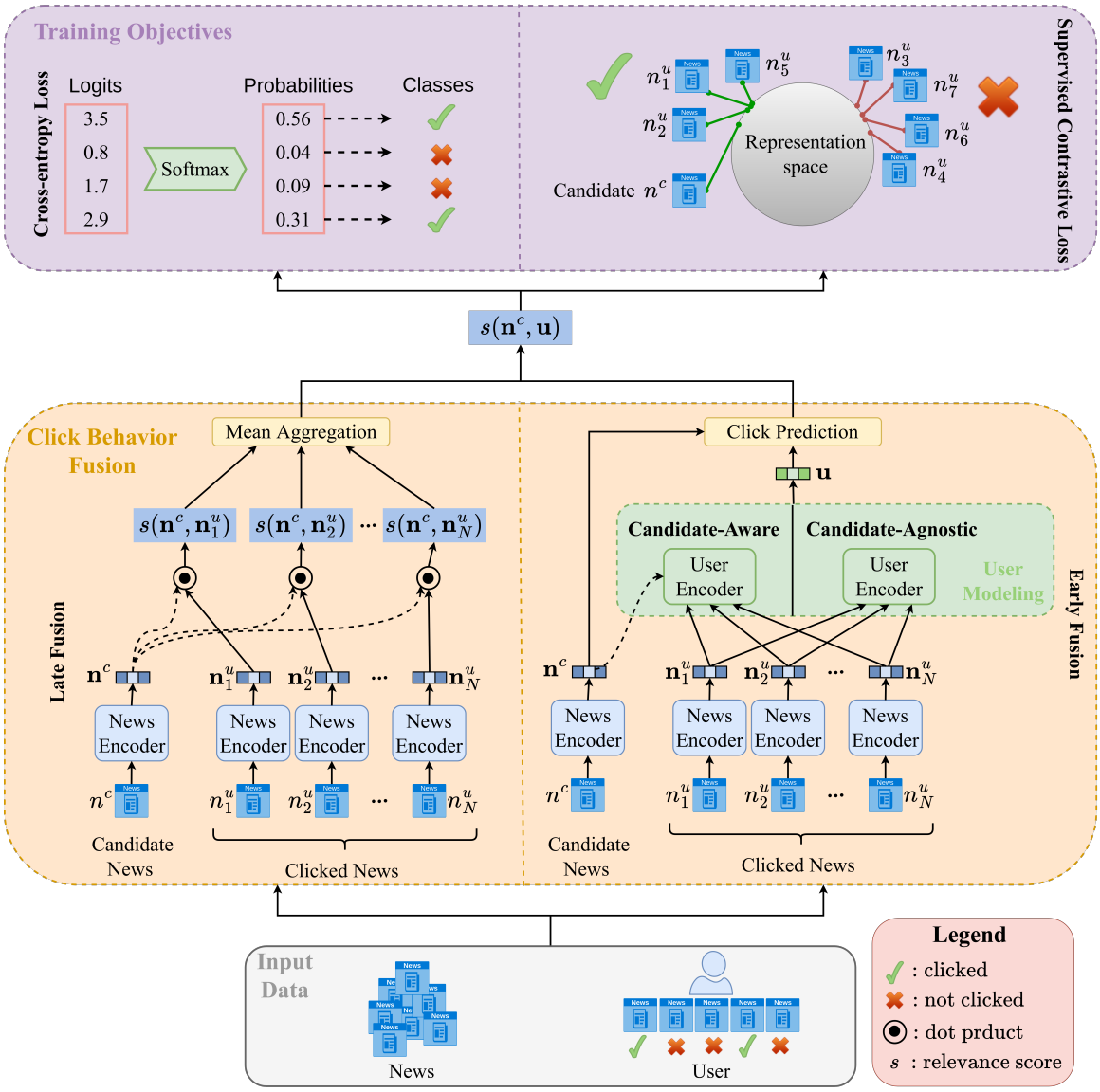


Figure 4.1: Illustration of the unified neural news recommendation framework, focusing on three crucial design dimensions: (i) candidate-awareness in user modeling (green box), (ii) click behavior fusion (orange box), and (iii) training objectives (purple box).

additive attention modules. Models with multiple feature vectors produce final news embeddings by simply concatenating them (LSTUR, CAUM), or by attending over them (NAML, MINS).

4.2.2 Click Behavior Fusion

We question whether the design and computational complexity of *early fusion* (EF), i.e., existence of dedicated user encoders in state-of-the-art NNRs, is justified. To this end, we propose, as a lightweight alternative, the *late fusion* (LF) approach that replaces the elaborate user encoders with mean-pooling of dot-product scores between the embedding of the candidate n^c and the embeddings of the clicked news n_i^u. Given a candidate news n^c and a sequence of news clicked by the user $\mathcal{H} = [n_1^u, ..., n_N^u]$, we compute the relevance score of the candidate news with regards to the user u's history as $s(\mathbf{n}^c, u) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{n}^c \cdot \mathbf{n}_i^u$, where $\mathbf{n}$ denotes the embedding of a news learned by the news encoder and N the history length.

Although LF suggests that explicitly encoding user behavior may not be necessary for click

prediction, user embeddings are still needed in collaborative-filtering models (Li and Wang, 2019). Note that the LF formulation above is equivalent to the dot product between the candidate embedding $\mathbf{n}^c$ and the mean of embeddings of the user's clicked news $\mathbf{n}_i^u$, $s(\mathbf{n}^c, u) = \mathbf{n}^c \cdot \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{n}_i^u\right)$. This means that LF can also seamlessly provide user embeddings (simply as averages of clicked news embeddings) if needed. LF can thus been seen as a *parameterless* user encoder, i.e., a computationally efficient alternative to complex parameterized UEs in existing EF models. Because (i) we produce embeddings of candidate and clicked news independently, and (ii) yield user embeddings as averages of clicked news embeddings, LF models are candidate agnostic (C-AG).

4.2.3 Training Objectives

The vast majority of existing work on neural news recommendation, regardless of the concrete user modeling architecture, tunes the parameters by minimizing the arguably most straightforward classification objective, cross-entropy loss (with negative sampling; see Fig. 4.1), and largely fails to explore effective alternatives, foremost contrastive objectives (Oord et al., 2018; Wu et al., 2022b). This prevents understanding of models effectiveness under different training regimes. We address this limitation by training all models (see § 4.2.1) not only with (1) common cross-entropy loss (with negative sampling), but also via (2) a contrastive learning objective, in particular supervised contrastive loss (Khosla et al., 2020).

4.3 Experimental Setup

Data. We conduct experiments on the MINDsmall and MINDlarge datasets (Wu et al., 2020e). For details on the dataset splits and corresponding statistics, we refer the reader to § 3.2.2.

Implementation and Optimization Details. We use 300-dimensional pretrained Glove embeddings (Pennington et al., 2014) and 100-dimensional TransE embeddings (Bordes et al., 2013) pretrained on Wikidata to initialize respectively the word and entity embeddings of the NNRs under comparison. We set the maximum history length to 50. Following Wu et al. (2022b), our negative sampling creates four negatives per positive example. We find the optimal temperature for SCL using the validation performance, sweeping the interval $[0.08, 0.3]$ with a 0.02 step. We train with batch size of 512 for all C-AG models, 256 for DKN and only 64 for CAUM (due to computational limitations). We set all other model-specific hyperparameters, to optimal values reported in the respective papers. We train all models with mixed precision, under a fixed computational budget: for 25 epochs on MINDsmall and 10 epochs on MINDlarge. We optimize with the Adam algorithm (Kingma and Ba, 2014), with the learning rate set to $1e$-4. We repeat each experiment five times (with different random seeds) and report averages (and std. deviation) for common metrics: AUC, MRR, NDCG@5, and NDCG@10. Each model is trained on a single NVIDIA Tesla V100 GPU with 32GB memory. Our implementation is publicly available.[2]

4.4 Results and Discussion

Table 4.1 shows the performance on MINDsmall and MINDlarge for both C-AG (NPA, NAML, NRMS, LSTUR, CenNewsRec, and MINS) and C-AW models (DKN, CAUM), under four different configurations of our comparative evaluation framework: (i) user modeling with EF versus LF, combined with (2) training with CE versus SCL objective. We next dissect the results along the three axes of our framework (Section 4.2): user modeling, click behavior fusion, and training objectives.

Candidate-Agnostic versus Candidate-Aware NNRs. We analyze C-AG versus C-AW models under their default EF configuration, since with LF all models become candidate-agnostic. CAUM, with the

[2]Code available at: https://github.com/andreeaiana/simplifying_nnr

		MINDsmall							
		AUC		MRR		NDCG@5		NDCG@10	
Model	**CBF**	CE	SCL	CE	SCL	CE	SCL	CE	SCL
NPA	EF	$54.72_{\pm 0.62}$	$56.48_{\pm 0.68}$	$29.04_{\pm 0.72}$	$28.38_{\pm 0.61}$	$26.91_{\pm 0.83}$	$26.58_{\pm 0.59}$	$33.19_{\pm 0.78}$	$32.63_{\pm 0.41}$
	LF	$55.15_{\pm 0.91}$	$57.32_{\pm 1.17}$	$28.58_{\pm 0.30}$	$27.49_{\pm 1.03}$	$26.40_{\pm 0.43}$	$25.52_{\pm 1.05}$	$32.87_{\pm 0.37}$	$31.83_{\pm 0.87}$
NAML	EF	$50.09_{\pm 0.02}$	$57.13_{\pm 1.08}$	$33.58_{\pm 0.53}$	$32.19_{\pm 0.66}$	$31.58_{\pm 0.61}$	$30.43_{\pm 0.71}$	$37.99_{\pm 0.49}$	$36.85_{\pm 0.66}$
	LF	$50.01_{\pm 0.00}$	$62.73_{\pm 0.51}$	$33.68_{\pm 0.77}$	$32.03_{\pm 0.66}$	$31.81_{\pm 0.89}$	$30.28_{\pm 0.65}$	$38.10_{\pm 0.76}$	$36.62_{\pm 0.56}$
NRMS	EF	$52.62_{\pm 1.32}$	$59.85_{\pm 0.63}$	$27.58_{\pm 0.76}$	$29.19_{\pm 0.71}$	$25.73_{\pm 0.48}$	$27.15_{\pm 0.85}$	$32.31_{\pm 0.49}$	$33.72_{\pm 0.67}$
	LF	$58.86_{\pm 1.02}$	$60.25_{\pm 1.12}$	$31.82_{\pm 0.67}$	$30.66_{\pm 0.58}$	$29.94_{\pm 0.65}$	$28.72_{\pm 0.58}$	$36.29_{\pm 0.58}$	$35.11_{\pm 0.59}$
LSTUR	EF_{ini}	$53.47_{\pm 1.20}$	$55.39_{\pm 0.47}$	$29.63_{\pm 0.45}$	$28.13_{\pm 0.66}$	$28.04_{\pm 0.45}$	$26.55_{\pm 0.68}$	$34.38_{\pm 0.45}$	$32.93_{\pm 0.60}$
	EF_{con}	$50.24_{\pm 0.03}$	$59.75_{\pm 1.39}$	$31.84_{\pm 0.69}$	$31.30_{\pm 0.81}$	$30.09_{\pm 0.78}$	$30.29_{\pm 1.30}$	$36.41_{\pm 0.67}$	$36.22_{\pm 0.72}$
	LF	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	$33.83_{\pm 0.59}$	$33.86_{\pm 0.64}$	$31.93_{\pm 0.67}$	$32.01_{\pm 0.66}$	$37.97_{\pm 0.64}$	$38.06_{\pm 0.58}$
CenNewsRec	EF	$53.96_{\pm 0.78}$	$59.97_{\pm 0.44}$	$28.30_{\pm 0.53}$	$30.59_{\pm 0.81}$	$26.52_{\pm 0.40}$	$28.50_{\pm 0.79}$	$32.89_{\pm 0.33}$	$34.77_{\pm 0.70}$
	LF	$59.32_{\pm 0.63}$	$61.93_{\pm 0.69}$	$32.84_{\pm 0.81}$	$32.22_{\pm 0.81}$	$30.90_{\pm 0.81}$	$30.37_{\pm 0.79}$	$37.15_{\pm 0.64}$	$36.59_{\pm 0.69}$
MINS	EF	$50.63_{\pm 0.34}$	$62.89_{\pm 1.67}$	$33.73_{\pm 1.04}$	$32.40_{\pm 0.35}$	$31.88_{\pm 1.05}$	$30.69_{\pm 0.36}$	$38.34_{\pm 0.94}$	$37.14_{\pm 0.34}$
	LF	$59.11_{\pm 1.22}$	$64.18_{\pm 0.69}$	$35.04_{\pm 0.54}$	$34.10_{\pm 0.63}$	$33.17_{\pm 0.64}$	$32.27_{\pm 0.71}$	$39.44_{\pm 0.56}$	$38.46_{\pm 0.57}$
DKN	EF	$50.00_{\pm 0.00}$	$51.05_{\pm 2.34}$	$26.36_{\pm 0.62}$	$25.94_{\pm 0.89}$	$24.45_{\pm 0.66}$	$23.92_{\pm 1.06}$	$30.99_{\pm 0.60}$	$30.52_{\pm 0.89}$
	LF	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	$27.45_{\pm 0.59}$	$26.36_{\pm 0.81}$	$24.98_{\pm 0.55}$	$24.01_{\pm 0.79}$	$31.74_{\pm 0.56}$	$30.83_{\pm 0.81}$
CAUM	EF	$61.40_{\pm 1.02}$	$63.19_{\pm 0.92}$	$33.84_{\pm 0.59}$	$33.66_{\pm 0.80}$	$31.96_{\pm 0.61}$	$31.82_{\pm 0.91}$	$38.41_{\pm 0.50}$	$38.20_{\pm 0.76}$
	LF	$62.37_{\pm 0.83}$	$63.55_{\pm 0.68}$	$33.69_{\pm 0.55}$	$33.68_{\pm 0.68}$	$31.79_{\pm 0.51}$	$31.83_{\pm 0.79}$	$38.17_{\pm 0.47}$	$38.01_{\pm 0.70}$
				MINDlarge					
NPA	EF	$56.76_{\pm 0.15}$	$58.14_{\pm 0.81}$	$31.44_{\pm 0.47}$	$30.04_{\pm 0.55}$	$29.47_{\pm 0.45}$	$27.74_{\pm 0.64}$	$35.85_{\pm 0.42}$	$34.35_{\pm 0.57}$
	LF	$61.18_{\pm 0.52}$	$58.69_{\pm 0.84}$	$32.13_{\pm 0.63}$	$28.34_{\pm 0.49}$	$30.24_{\pm 0.68}$	$26.02_{\pm 0.68}$	$36.58_{\pm 0.57}$	$32.68_{\pm 0.67}$
NAML	EF	$50.15_{\pm 0.04}$	$60.40_{\pm 0.65}$	$33.20_{\pm 0.41}$	$34.17_{\pm 0.41}$	$31.32_{\pm 0.46}$	$32.47_{\pm 0.34}$	$37.89_{\pm 0.39}$	$38.91_{\pm 0.30}$
	LF	$50.00_{\pm 0.00}$	$65.38_{\pm 0.51}$	$32.71_{\pm 0.45}$	$33.46_{\pm 0.48}$	$31.03_{\pm 0.48}$	$31.73_{\pm 0.52}$	$37.60_{\pm 0.39}$	$38.21_{\pm 0.40}$
NRMS	EF	$54.56_{\pm 1.44}$	$62.76_{\pm 0.70}$	$31.94_{\pm 1.00}$	$32.45_{\pm 0.53}$	$30.04_{\pm 1.06}$	$30.53_{\pm 0.67}$	$36.62_{\pm 0.96}$	$36.87_{\pm 0.56}$
	LF	$56.10_{\pm 2.11}$	$63.60_{\pm 1.06}$	$32.92_{\pm 0.69}$	$32.43_{\pm 0.62}$	$31.67_{\pm 1.07}$	$30.61_{\pm 0.80}$	$37.80_{\pm 0.35}$	$37.11_{\pm 0.68}$
LSTUR	EF_{ini}	$50.03_{\pm 0.06}$	$56.89_{\pm 1.50}$	$32.45_{\pm 2.36}$	$31.27_{\pm 1.54}$	$30.93_{\pm 2.40}$	$29.80_{\pm 1.80}$	$37.38_{\pm 2.41}$	$36.24_{\pm 1.56}$
	EF_{con}	$51.37_{\pm 0.44}$	$54.26_{\pm 0.40}$	$27.66_{\pm 0.45}$	$26.48_{\pm 0.22}$	$25.88_{\pm 0.49}$	$24.65_{\pm 0.24}$	$32.34_{\pm 0.49}$	$31.13_{\pm 0.24}$
	LF	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	$34.73_{\pm 0.61}$	$33.05_{\pm 0.25}$	$33.23_{\pm 0.58}$	$31.61_{\pm 0.39}$	$39.19_{\pm 0.54}$	$37.70_{\pm 0.29}$
CenNewsRec	EF	$56.42_{\pm 0.80}$	$64.52_{\pm 0.41}$	$33.66_{\pm 0.32}$	$33.54_{\pm 0.42}$	$31.94_{\pm 0.27}$	$31.79_{\pm 0.46}$	$38.27_{\pm 0.22}$	$38.15_{\pm 0.42}$
	LF	$53.32_{\pm 0.67}$	$64.19_{\pm 0.58}$	$33.22_{\pm 0.44}$	$33.35_{\pm 0.3}$	$31.43_{\pm 0.48}$	$31.69_{\pm 0.43}$	$37.88_{\pm 0.39}$	$38.08_{\pm 0.36}$
MINS	EF	$51.75_{\pm 0.15}$	$65.85_{\pm 0.54}$	$34.34_{\pm 0.20}$	$34.39_{\pm 0.46}$	$32.45_{\pm 0.42}$	$32.61_{\pm 0.50}$	$39.07_{\pm 0.38}$	$39.12_{\pm 0.46}$
	LF	$53.77_{\pm 0.59}$	$66.69_{\pm 0.79}$	$34.88_{\pm 0.18}$	$34.82_{\pm 0.69}$	$32.98_{\pm 0.16}$	$33.41_{\pm 0.70}$	$39.53_{\pm 0.16}$	$39.57_{\pm 0.62}$
DKN	EF	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	$25.24_{\pm 0.41}$	$24.80_{\pm 0.28}$	$23.36_{\pm 0.72}$	$22.60_{\pm 0.29}$	$29.96_{\pm 0.46}$	$29.08_{\pm 0.32}$
	LF	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	$29.12_{\pm 0.42}$	$27.84_{\pm 1.12}$	$26.32_{\pm 0.33}$	$25.45_{\pm 1.02}$	$33.24_{\pm 0.41}$	$32.14_{\pm 0.95}$
CAUM	EF	$67.09_{\pm 0.83}$	$66.44_{\pm 0.90}$	$35.32_{\pm 0.48}$	$35.13_{\pm 0.51}$	$33.63_{\pm 0.60}$	$33.41_{\pm 0.58}$	$40.06_{\pm 0.50}$	$39.86_{\pm 0.52}$
	LF	$53.13_{\pm 0.27}$	$65.89_{\pm 0.17}$	$34.46_{\pm 0.35}$	$34.46_{\pm 0.12}$	$32.62_{\pm 0.34}$	$32.85_{\pm 0.13}$	$39.17_{\pm 0.28}$	$39.26_{\pm 0.07}$

Table 4.1: Recommendation performance of the compared models under combinations of click behavior fusion (CBF) and training objectives. We report averages and standard deviations across five different runs.

most complex and candidate-aware UE, generally outperforms all other models under both training regimes (CE and SCL) and for most evaluation metrics. The gaps are particularly prominent on the large training dataset, MINDlarge, w.r.t. the AUC metric. This result alone could mislead to a conclusion that more complex, candidate-aware user modeling is necessary for better recommendation. The fact that (1) DKN, as the other C-AW model in our evaluation – generally performs much worse than C-AG models, as well as that (2) our LF models variants with trivial, parameterless UEs match or surpass the performance of CAUM with EF, undermine this conclusion. With the exception of DKN, all other models exhibit better performance when trained on the larger MINDlarge dataset. NAML and MINS are, however, competitive (except w.r.t. AUC metric) on MINDsmall, but fall behind CAUM on MINDlarge, suggesting that CAUM's elaborate UE benefits the most from more training data.

One confounding factor that we do not control for, however, and which warrants a mindful comparison of the results, is that models differ not just in UE, but also in NE components, i.e., w.r.t. how they encode news and which features they use as input. For example, NAML and MINS, with an identical NE, achieve similar performance on MINDsmall. On MINDlarge, however, the more complex UE of MINS brings substantial gains over the simpler UE of NAML (but only under standard EF fusion and CE training).

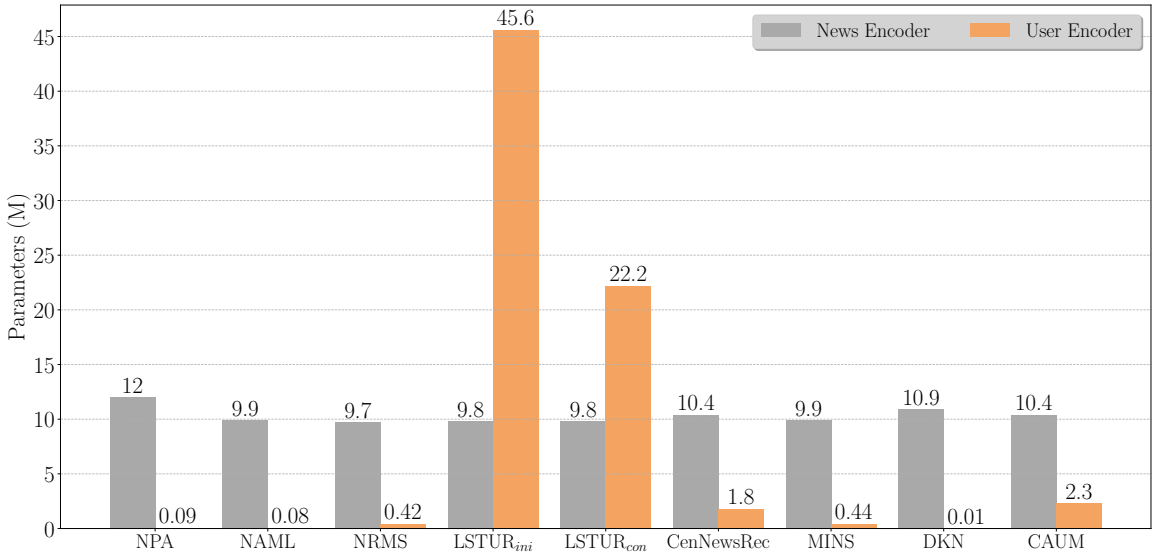


Figure 4.2: Number of model parameters (in millions) on MINDsmall.

Early versus Late Click Behavior Fusion. Replacing complex EF-based UEs with the simple parameterless LF that we propose brings substantial performance gains across the board. Averaged across all models and both training objectives, LF brings massive gains of 5.58 and 4.63 MRR points on MINDsmall and MINDlarge, respectively. Equally importantly, with LF – i.e., with the same parameterless UE – models exhibit mutually much more similar performance than under EF, with other models generally closing the gap to CAUM. This suggests that LF makes differences in NEs architectures across models less consequential, thus not only simplifying UE with parameterless averaging of clicked news embeddings, but also allowing for simpler news encoders.

Cross-Entropy versus Supervised Contrastive Loss. Overall, we find SCL to be a viable alternative to the common cross-entropy based classification with negative sampling (compare columns CE and SCL across evaluation metrics in Table 4.1). SCL brings large gains over CE in terms of AUC (+8.26 points on MINDsmall and +12.14 points on MINDlarge, averaged across all models, in both EF and LF variants). This suggests that, SCL leads to better separation of clicked and not clicked news in the representation space. In contrast, SCL falls slightly behind CE according to ranking measures, MRR and NDCG (-1.54 and -1.78 MRR points on MINDsmall and MINDlarge, respectively). We hypothesize that this is because of hard negatives – news not clicked by the user that resemble user's clicked news – for which CE more directly signals irrelevance: these likely become highly-ranked false positives for SCL-trained models.

Model Size. Finally, we quantify the reduction in model parameters that LF brings w.r.t. EF. Fig. 4.2 shows the number of trainable parameters in original EF configurations, on MINDsmall.[3][4] While the NE accounts for the majority of parameters in most models, the plot shows that the proportion of UE parameters is non-negligible for several models, and largest by a wide margin for LSTUR. With a parameterless UE, along with performance gains, LF brings a relative reduction of model size of 14.7%, 18.1%, and massive 82.3% for CenNewsRec, CAUM, and LSTUR$_{\text{ini}}$, respectively.

[3]For some models, e.g., LSTUR with its user embedding matrix, the number of parameters depends on the size of the training data.

[4]We report similar results on MINDlarge in Appendix D.

4.5 Conclusion

Rapid development of personalized neural news recommenders hinders fair comparative model evaluations and systematic analyses of design choices. In this Chapter, we introduced a unified framework for neural news recommendation focusing on three crucial design dimensions of NNRs: (i) candidate-awareness in user modeling, (ii) click behavior fusion, and (iii) training objectives. Extensive evaluation of a wide range of recent state-of-the-art models reveals that NNRs can be drastically simplified: replacing complex user encoders with parameterless aggregation of clicked news embeddings brings substantial performance gains across the board, reducing at the same time model complexity. Further, we show that contrastive learning can be a viable alternative to standard classification-based (cross-entropy) loss. We hope that our findings will inspire more transparent evaluation, including systematic model ablations to uncover the components that drive performance. In the next Chapter, we will extend out analysis of common components of NNRs to the main drivers of news and user modeling modules, namely their underlying encoder architectures.

CHAPTER 5

ENCODER SELECTION

A system is never the sum of its parts; it's the product of their interaction.

RUSSELL L. ACKOFF

* In the previous Chapter, we showcased that increased architectural complexity in user modeling does not necessarily translate into superior recommendation performance. We further demonstrated that different training objectives are suited to different optimization goals, highlighting the need to deliberately select objectives based on the intended outcome (e.g., accurate click prediction versus effective candidate ranking). However, user modeling and training objectives constitute only a subset of the components that jointly determine the behavior and performance of a neural recommender.

Encoder architectures play a pivotal role in neural news recommenders by embedding the semantic and contextual information of news and users (cf. § 3.1.1 and § 3.1.2). As a result, a substantial body of work has focused on enhancing the representational capabilities of news and user encoders to improve recommender performance. Despite their central role in determining the quality of news and user representations, existing analyses of encoder designs predominantly assess their impact on the overall downstream recommendation performance or by studying components in isolation. Such evaluations provide a limited, one-sided view of encoder behavior, overlooking similarities and differences in the representations they induce and their interactions within the system, which can ultimately lead to sub-optimal model selection (**C1**).

In the following, we perform a comprehensive analysis of encoder architectures in neural news recommender systems. We systematically evaluate the most prominent news and user encoder architectures, focusing on their (i) representational similarity, measured with the Central Kernel Alignment, (ii) overlap of generated recommendation lists, quantified with the Jaccard similarity, and (iii) the overall recommendation performance. Our analysis reveals that the complexity of certain encoding techniques is often empirically unjustified, highlighting the potential for simpler, more efficient architectures. By isolating the effects of individual components, we provide valuable insights for researchers and practitioners to make better informed decisions about encoder selection and avoid unnecessary complexity in the design of news recommenders.

5.1 Introduction

Content-based neural models have become the state of the art in news recommendation. As discussed in Section 3.1, NNRs typically comprise a news encoder and a user encoder. The news encoder learns semantically meaningful representations of news articles, whereas the user encoder embeds

*This Chapter is adapted from: **Andreea Iana**, Goran Glavaš, and Heiko Paulheim. 2024. Peeling Back the Layers: An In-depth Evaluation of Encoder Architectures in Neural News Recommenders. In *INRA 2024: Proceedings of the International Workshop on News Recommendation and Analytics co-located with the 2024 ACM Conference on Recommender Systems (RecSys 2024)*, volume Vol-3929, pages 1–15, Aachen, Germany. RWTH Aachen.

the preferences of users based on their click history (Wu et al., 2023). NNRs take the candidate news articles and a user's reading history as input. The relevance of the candidate to the user is determined by comparing, with a scoring function, the latent representations of the two inputs, generated with the corresponding encoders. Given the key role of encoders in NNRs, a significant body of research has focused on improving the quality of news encoding and user modeling to improve recommendation performance (Karimi et al., 2018; Raza and Ding, 2022; Wu et al., 2023).

On the one hand, ablation studies of recommenders typically analyze individual model components in isolation, neglecting other architecturally comparable model designs (Wu et al., 2019a,e; An et al., 2019). At the same time, we see emerging evidence that widely used NNRs exhibit similar performance despite varying model complexities, and that the overall complexity of the recommenders' architecture could be reduced (Möller and Padó, 2024, Chapter 4). This highlights the need for a more granular comparison of the individual building blocks to understand their behavior and impact on the overall system. While we in Chapter 4 and Möller and Padó (2022) evaluated NNR components such as scoring functions and training objectives, a systematic analysis of encoder architectures is still lacking. Such insights would enable researchers and practitioners alike to make more informed choices about encoder selection in NNR design.

On the other hand, progress in the architectural design of news and user encoders is generally measured in terms of the recommender's overall classification and ranking capability (Wu et al., 2019a; An et al., 2019; Qi et al., 2022b; Wu et al., 2021b, Chapter 4). Nonetheless, the quality of the embeddings produced by the news and user encoders is equally crucial, given the reliance of the recommender on the dense retrieval paradigm. Therefore, evaluating NNRs and their components solely in terms of downstream recommendation performance provides a simplified perspective, potentially overlooking subtle differences in the encoders' behavior. We thus argue that investigating the similarity of embeddings generated by various news and user encoders would offer a more nuanced understanding of their behavior, in turn benefiting the model selection process.

In this Chapter, we perform a systematic analysis of the encoder architectures of NNRs. Unlike conventional evaluation studies, we isolate the effects of each core component to the largest possible extent. Concretely, we analyze the most prominent news and user encoder architectures in terms of (i) the similarity of learned news, and respectively, user representations, using the Central Kernel Alignment (Kornblith et al., 2019) metric, (ii) the similarity of the generated recommendation lists, quantified by means of the Jaccard coefficient, and (iii) the impact on the overall recommendation performance. Our findings provide a better understanding of encoder architectures in news recommenders, not only from a recommendation performance perspective, but also in terms of their representational similarity. We demonstrate that the complexity of some encoding techniques is often empirically unjustified, emphasizing the potential benefits of simpler, more efficient architectures. These results fundamentally challenge the common practice of over-engineering NNR encoders. Consequently, we derive three key takeaways, arguing that (1) the semantic accuracy of representations produced by news encoders is crucial for effective recommendation, that (2) user encoders can be significantly simplified without sacrificing performance, and lastly, (3) we advocate for more rigorous evaluation to guide better informed model selection.

5.2 Related Work

Neural news recommenders have significantly advanced in recent years, with encoder architectures playing a key role in capturing the semantic and contextual information of news articles and user profiles. Consequently, a large strand of work has focused on improving the representational capabilities of recommenders by developing ever more accurate, and often complex, news encoding and user modeling architectures. As such, these works have analyzed individual aspects of the NNR components, such as the use of different attention mechanisms in the news or user encoder (Wu et al., 2019a,e; Qi et al., 2020), the impact of various user modeling (An et al., 2019; Qi et al., 2022b;

Wang et al., 2022c; Qi et al., 2020, Chapter 4) or news embedding (Wang et al., 2018b; Wu et al., 2019a,e; Qi et al., 2020; Wu et al., 2021b; Li et al., 2022, Chapter 9) techniques, or the importance of modeling different news features (Wang et al., 2018b; Wu et al., 2019a,c, 2020c; Xun et al., 2021; Qi et al., 2021c; Wu et al., 2021c) and user characteristics (Wu et al., 2019e; An et al., 2019; Wu et al., 2019d, 2022g). Ablation studies in these cases are usually conducted in isolation for the component under consideration, without taking into account the broader architectural context.

In contrast, another strand of work has started evaluating the impact of NNR components or training strategies across an array of recommendation approaches. For example, Wu et al. (2021b) have investigated the usage of various pretrained language models as the backbone of widely used NNRs. Möller and Padó (2022) have evaluated the impact of scoring functions, whereas in Chapter 4 we have analyzed different user modeling techniques and training objectives. We have highlighted the similar recommendation performance achieved by certain models despite differences in architectures and complexity, emphasizing the potential to simplify the design of news recommender systems. While these works shed new light on core components of the recommendation model, their evaluation is most often solely based on the downstream recommendation performance.

The similarity of encoders in NNRs can additionally be measured in terms of their generated representations. More generally, there exist numerous methods for quantifying the similarity of neural networks. Two main categories include (i) representational similarity, which assesses differences in the activations of intermediate layers of neural networks, and (ii) functional similarity, which compares the networks' outputs in relation to their task (Klabunde et al., 2025). Several works have focused on evaluating the representational similarity of (large) language models (Wu et al., 2020f; Klabunde et al., 2023; Brown et al., 2023; Freestone and Santu, 2024) or of embedding models in Retrieval Augmented Generation systems (Caspari et al., 2024), which are often employed as the news encoding component of NNRs.

Nevertheless, to the best of our knowledge, no work so far compares neither user encoders nor news encoders with respect to representational and functional similarity. In what follows, we fill this gap by comprehensively analyzing the primary components of NNRs' encoder architectures for both news and user inputs.

5.3 Methodology

We firstly introduce the building blocks of personalized NNRs. Afterwards, we discuss metrics to evaluate both the recommendation performance, as well as the representational similarity of the news and user encoders.

5.3.1 Encoders of Neural News Recommenders

The building blocks of neural news recommender systems (i.e., news encoder, user encoder, scoring function, training objective) altogether drive the overall performance of the recommender. Since the news and user encoders determine what information of the documents and users is embedded by the model, and ultimately, propagated through the recommendation pipeline, both types of encoders play a similarly important role in model selection. We introduce the abbreviations used for the remainder of the Chapter in Table 5.1.

News Encoder Architectures. The news encoder can generally be decomposed into a *text encoder*, which embeds the textual content of a news article, and several *feature-specific encoders* (e.g., category, sentiment, entity encoder), which learn to represent further input features different from text chunks. While the former represents a key component of all NNRs, the latter types of encoders are optional and only utilized whenever the textual content is enriched with additional features which might capture or emphasize other aspects of a news article. Lastly, the news encoder combines

Abbreviation	Description
Att	attention network
AddAtt	additive attention (Bahdanau et al., 2015)
MHSA	multi-head self-attention (Vaswani et al., 2017)
$\text{PLM}_{\text{[CLS]}}$	the PLM's output `[CLS]` token representation
$\text{PLM}_{\text{tokenemb+Att}}$	PLM's token embeddings pooled with an attention network (Wu et al., 2021b)
Con	concatenation
Linear	linear layer
GRU	gated recurrent unit (Cho et al., 2014)
CandAware	candidate-aware user encoder (Qi et al., 2022b)

Table 5.1: Abbreviations and their description.

Text Embedding Type	Text Encoder	References
word embeddings	`CNN + AddAtt`	Wu et al. (2019a); An et al. (2019); Wu et al. (2019c,d); Sheu and Li (2020); Santosh et al. (2020)
	`MHSA + AddAtt`	Gao et al. (2018a); Wu et al. (2019e, 2020c,d); Ge et al. (2020); Tran et al. (2021); Qi et al. (2021c); Wu et al. (2022d, 2021a,d); Qi et al. (2022b); Wang et al. (2022c)
	`CNN + MHSA + AddAtt`	Qi et al. (2020)
language model	$\text{PLM}_{\text{tokenemb+Att}}$	Wu et al. (2020b, 2021b); Zhang et al. (2021a,b)
	$\text{PLM}_{\text{[CLS]}}$	Jia et al. (2021); Wu et al. (2022f); Li et al. (2022); Shivaram et al. (2022), Chapter 7
	`SE`	Chapter 9

Table 5.2: Text encoder architectures.

the intermediate embeddings produced by the text and feature-specific encoders into a news-level representation by means of a *multi-feature aggregation strategy*.

We list the most used types of text encoders that we consider in our analysis in Table 5.2, alongside examples of NNRs using them. We distinguish between text encoders that rely on pretrained word embeddings, contextualized by means of convolutional or self-attention networks, and the more recent architectures that employ pretrained language models.[1] We additionally consider the most common multi-feature aggregation approaches used to integrate text and other content feature (e.g., category) embeddings into the unified news representation, as shown in Table 5.3.

User Encoder Architectures. Parameterized user encoders represent the most popular user modeling technique. They learn user representations by means of sequential or attentive networks that contextualize the embeddings of clicked news based on patterns in the user's click behavior. User encoders can be further differentiated into candidate-agnostic (i.e., users are encoded separately from candidate news) and candidate-aware (i.e., the user-level aggregation contextualizes the embeddings of clicked news against the embedding of each candidate) encoders, as discussed in Chapter 4. In the previous Chapter 4, we proposed the parameter-free late fusion (`LF`) approach. `LF` first averages the clicked news embeddings $\mathbf{n}_i^u$ to a user embedding $\frac{1}{N}\sum_{i=1}^{N}\mathbf{n}_i^u = \mathbf{u}$. The relevance score is then

[1]Note that in this Chapter we do not evaluate encoders which rely on news or user graphs, as such graphs are heavily dataset-dependent. We instead focus on the most used core components of encoders, and leave the analysis of graph-based techniques for future work.

Multi-feature aggregation	References
`AddAtt`	Wu et al. (2019a); Santosh et al. (2020); Wu et al. (2021a); Sun et al. (2021); Raza and Ding (2021); Wang et al. (2022c)
`Linear`	Tran et al. (2021); Qi et al. (2022b)
`Con`	An et al. (2019); Ge et al. (2020); Han et al. (2021)

Table 5.3: Multi-feature aggregation strategies for combining textual and categorical representations of news.

User Encoder	References
`LF`	Chapters 7 and 9
`AddAtt`	Gao et al. (2018a); Wu et al. (2019a,d,c); Liu et al. (2020a); Santosh et al. (2020); Han et al. (2021); Zhang et al. (2021c)
`MHSA+AddAtt`	Wu et al. (2019e, 2020c,b, 2022d); Xun et al. (2021); Wu et al. (2021d)
$\texttt{GRU}_{\texttt{ini}}$	An et al. (2019); Sun et al. (2021)
$\texttt{GRU}_{\texttt{con}}$	An et al. (2019); Tran et al. (2021)
`GRU+MHSA+AddAtt`	Qi et al. (2020)
`CandAware (CNN+MHSA+AddAtt)`	Qi et al. (2022b)

Table 5.4: User encoder architectures.

obtained using the inner product between the candidate news embedding $\mathbf{n}^c$ and the user embedding $\mathbf{u}$. Table 5.4 lists the main user encoder architectures that we evaluate, together with examples of models using them.

5.3.2 Similarity Evaluation

We evaluate news and user encoders on three dimensions: (i) downstream recommendation performance, (ii) similarity of generated recommendations, and (iii) similarity of learned news or user representations.

Downstream Recommendation Performance. NNRs are usually evaluated with regards to classification (e.g., AUC) and ranking (e.g., MRR, NDCG) performance. In this work, we focus on the ranking performance, which we quantify using NDCG@k.

Similarity of Generated Recommendations. We analyze the retrieval similarity of recommenders that use different news or user encoder architectures by the similarity of their top-k recommended articles. Specifically, for the same set of users, we firstly generate the corresponding recommendation lists $\mathcal{R}$ and $\mathcal{R}'$ with models M and M', respectively. We then measure the similarity of retrieved results with the Jaccard similarity coefficient:

$$Jaccard(\mathcal{R}, \mathcal{R}') = \frac{|\mathcal{R} \bigcap \mathcal{R}'|}{|\mathcal{R} \bigcup \mathcal{R}'|} \tag{5.1}$$

where $|\mathcal{R} \bigcap \mathcal{R}'|$ denotes the set of articles recommended by both models, and $|\mathcal{R} \bigcup \mathcal{R}'|$ the union of all unique news recommended by the two models. The Jaccard similarity score is bounded in the $[0, 1]$ interval, with 1 indicating that both models recommend an identical set of news. Note that the lengths of both recommendation lists will be equal to the full set of candidate news $\mathcal{N}_u^c$ for a given user u, namely $|\mathcal{R}| = |\mathcal{R}'| = |\mathcal{N}_u^c|$, regardless of the recommendation model used. Thus, to differentiate the retrieval performance of two models, we compute the Jaccard similarity only

for the top-k recommendations, ordered descendingly by the recommendation scores. Note that in comparison to NDCG@k, the Jaccard similarity measures the overlap of the recommended news between two models without considering the order of the articles in the recommendation set.

Embedding Similarity. Numerous measures quantify the representational similarity of neural networks (Klabunde et al., 2025). Many of these methods require an identical dimensionality of the compared embeddings or an alignment of the latent representation spaces across models. Since these constraints are not straightforwardly met by the embeddings produced with different news and user encoder architectures, we choose to measure the similarity of embeddings using the Centered Kernel Alignment (CKA) with a linear kernel (Kornblith et al., 2019). Concretely, for a given representation $\mathbf{E}$, we firstly mean-center it column-wise. Afterwards, we compute the pair-wise similarity of the representation of each instance i to all other instances in $\mathbf{E}$. Each row i in the resulting similarity matrix $\mathbf{S}$ thus comprises the similarity between instance's i embedding and all other embeddings, including itself. For two different models with the same number of embeddings $\mathbf{E}$ and $\mathbf{E}'$, the resulting representational similarity matrices $\mathbf{S}$ and $\mathbf{S}'$, respectively, can be directly compared using the Hilbert-Schmidt Independence Criterion (HSIC, Gretton et al., 2005) as follows:

$$CKA(\mathbf{E}, \mathbf{E}') = \frac{HSIC(\mathbf{S}, \mathbf{S}')}{\sqrt{HSIC(\mathbf{S}, \mathbf{S})HSIC(\mathbf{S}', \mathbf{S}')}} \tag{5.2}$$

CKA similarity scores are bounded to $[0, 1]$, with a score of 1 denoting equivalent representations.

5.4 Experimental Setup

Data. We conduct experiments on the MINDsmall (Wu et al., 2020e) dataset. For details on the dataset splits and corresponding statistics, we refer the reader to § 3.2.2.

Evaluation Setup. We separately evaluate the encoder architectures of the neural recommendation models. In all experiments, we consider both mono-feature (e.g., title) and multi-feature (e.g., title and categories) inputs for the news encoder. In the latter case, we learn category representations by means of a linear encoder that combines a category ID embedding layer with a dense layer (Wu et al., 2019a; An et al., 2019; Qi et al., 2022b; Wang et al., 2022c). Moreover, in our analysis of news encoder architectures, we adopt the *late fusion* approach (see Chapter 4) instead of the traditional parameterized user encoders. This evaluation setup allows us to isolate the effects of news enoders, and to avoid additional confounding factors stemming from the user encoder, which also influence the output of the NNR. Similarly, when evaluating the similarity of user encoder architectures, we keep the underlying news encoder of the recommender fixed, i.e., we analyze different user encoders for the same base news encoder.

Implementation and Optimization Details. We train all models with the standard cross-entropy loss, using dot product as the scoring function. We use 300-dimensional pretrained Glove embeddings (Pennington et al., 2014) to initialize the word embeddings of the word embedding-based text encoders. Additionally, we use $\text{RoBERTa}_{\text{Base}}$ (Liu et al., 2019b) and the news-specialized multilingual sentence encoder NaSE (see Chapter 9) for the PLM-based and SE-based text encoders, respectively. We fine-tune only the last four layers of the language models. Following prior work (Wu et al., 2022b), we sample four negatives per positive example during training. We set the maximum history length to 50, and train all models with mixed precision, the Adam optimizer (Kingma and Ba, 2014), and a batch size of 8. We train all NNRs with word embedding-based news encoders for 20 epochs, and those with language model-based news encoders for 10 epochs. We tune the main hyperparameters of all NNRs using grid search. Concretely, we search for the optimal learning rate in $\{1e{-}3, 1e{-}4, 1e{-}5\}$. We optimize the number of heads in the multi-head self-attention networks in $[8, 12, 16, 20, 24, 32]$, and the query vector dimensionality by sweeping the interval $[50, 200]$ with a step of 50. We run all

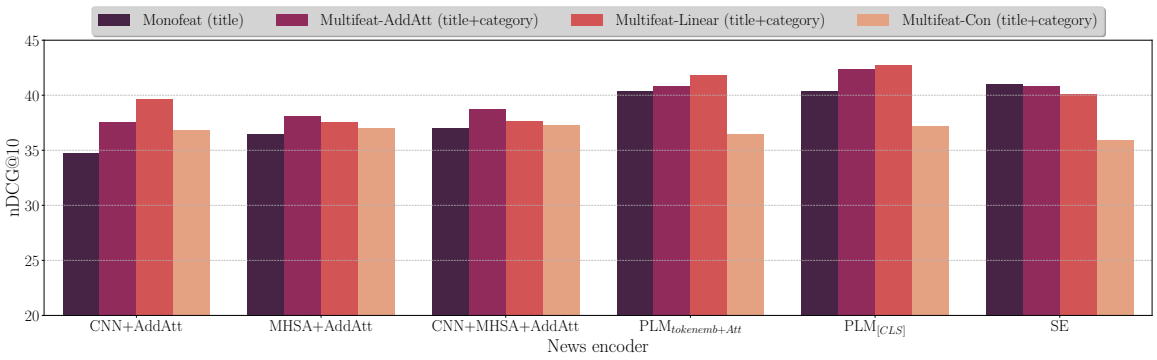


Figure 5.1: Ranking performance (nDCG@10) of recommenders depending on the news encoder architecture and input features.

experiments using the implementations available in the NewsRecLib library (see § 3.3.2), on a cluster with virtual machines, training each model on a single NVIDIA A100 40GB GPU.

5.5 Results and Discussion

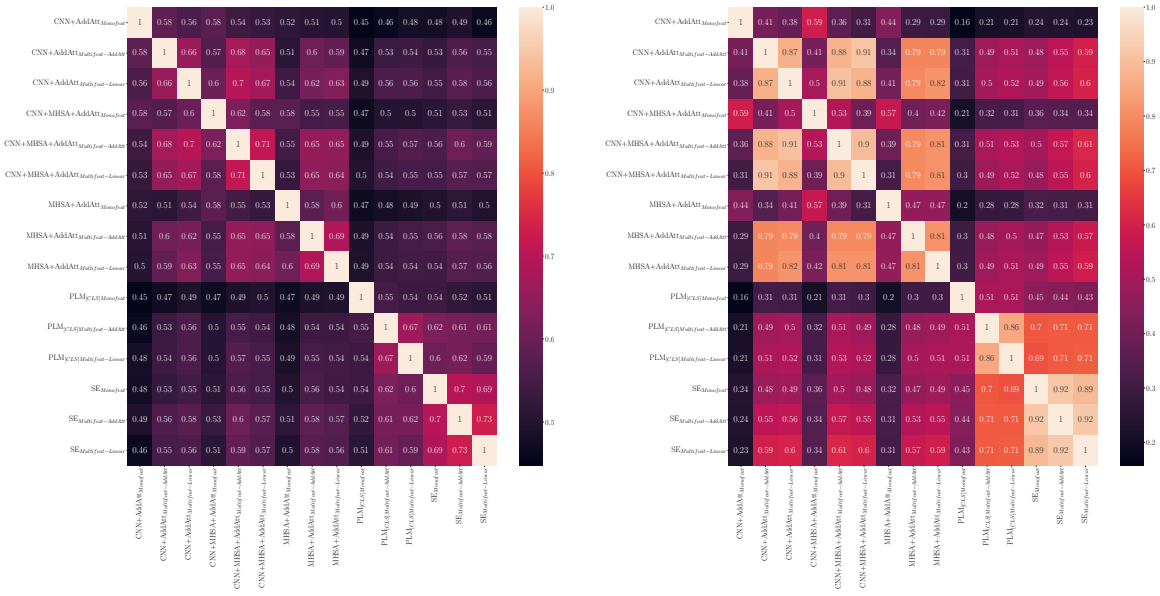
We begin by analyzing the similarity of core news encoder architectures, followed by an evaluation of user encoder similarity using the same base news encoding approach. In both cases, we first compare the architectures in terms of ranking performance and retrieval similarity, as these are standard evaluation approaches in the recommender systems field. We then assess the architectures from the perspective of pair-wise embedding similarity.

5.5.1 News Encoder Architectures

Fig. 5.1 shows the ranking performance, in terms of nDCG@10, of NNRs for different news encoders and input features. For the same input type, e.g. mono-feature, we find a high similarity between the performance of recommenders based on the same family of text encoders. Specifically, text encoders using pretrained static word embeddings are outperformed by those based on PLMs. Moreover, `MHSA+AddAtt` and `CNN+MHSA+AddAtt` appear to have nearly identical performance, despite the increased complexity of the latter architecture. Similarly, simply using the $[CLS]$ token representation produced by the PLM instead of pooling tokens with an attention network as proposed by Wu et al. (2021b) leads to slightly better performance while maintaining a lighter text encoder.

Our findings show that among the three multi-feature aggregation strategies, the `Linear` and `AddAtt` approaches always outperform the `Con` technique. This is intuitive, as the concatenation of vectors with varying dimensionality from non-aligned representation spaces will be sub-optimal. In contrast, both other aggregation strategies project the intermediate text and category embeddings in the same latent representation space. Most importantly, we find that leveraging categories in addition to textual news content as input features is most beneficial for word embedding-based text encoders, and becomes irrelevant or slightly detrimental for the domain-adapted sentence encoder. This can be explained, on the one hand, by the better representational capabilities of the much larger language models which acquire contextual understanding during pretraining compared to static word embeddings. On the other hand, sentence encoders, especially domain-specialized models such as NaSE (see Chapter 9), better capture nuances and topics from text due to their pretraining objectives that focus on the overall sentence-level semantics.

We find these similarities in ranking performance between the various news encoding architectures to be reflected in the similarity of retrieved articles. Fig. 5.2a illustrates the pair-wise Jaccard similarity scores between the top-10 recommended news per model. Note that we exclude `PLM`$_{\text{tokenemb+Att}}$, as



(a) Jaccard similarity of top-10 recommended news.

(b) CKA similarity of news embeddings.

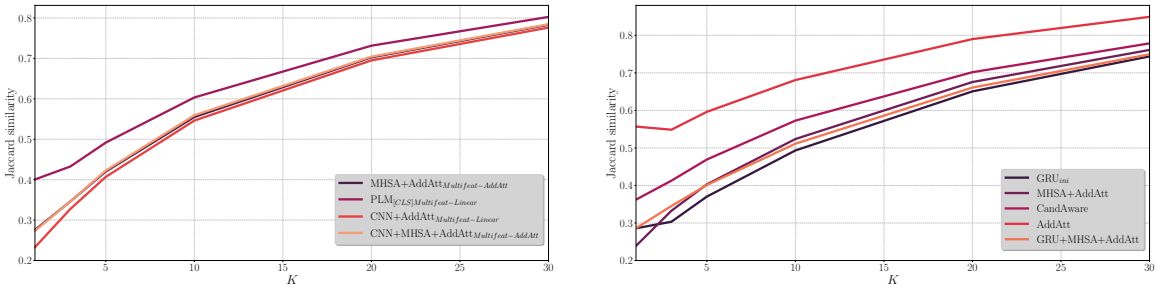
Figure 5.2: Retrieval and representational similarity of models with different news encoder architectures and input features. Each model's subscript indicates the type of input, and the multi-feature aggregation strategy, if used.

well as the Con multi-feature aggregation strategy from further analysis for the sake of brevity, and due to their poorer performance. As expected, models from the same family of text encoders show higher similarity scores. The lower Jaccard similarities across word embedding and PLM-based intra-family models using mono-feature versus multi-feature input supports our previous observation regarding the low relevance of categorical input for the domain-adapted SE.

The overall pair-wise Jaccard similarities could initially suggest that most news encoders result in little overlap in their recommendation lists. However, a Jaccard similarity score of 0.54 between two models for a list of $k = 10$ recommended items means that, in practice, the two models output 7 identical articles. Analogously, a score of 0.66 indicates an overlap of 8 out of 10 recommendations. As Fig. 5.2a shows, the recommendations generated by the various news encoder architectures differ by more than 3 articles in a list of length 10 only in rare cases. In other words, regardless of the architectural differences and complexities, the encoders retrieve, on average, the same articles in over 70% of the time.

Taking a look at the CKA similarity of the test set news embeddings produced with the different news encoders, shown in Fig. 5.2b, corroborates our hypothesis: intra-family news encoders tend to produce similar embeddings when using the same type of input features. The news-adapted SE constitutes the only exception, as its embeddings are not significantly influenced by leveraging categories as additional input features. Additionally, we observe a higher representational similarity between the CNN+AddAtt, MHSA+AddAtt, and CNN+MHSA+AddAtt models with multi-feature input, and a slightly lower similarity between $\text{PLM}_{[\text{CLS}]}$ and SE-based models. Overall, the high similarity of representations, of recommendation performance, and the large overlap of generated recommendations by the CNN+AddAtt, MHSA+AddAtt, and CNN+MHSA+AddAtt multi-feature news encoders contest the empirical contribution of incremental architectural changes in the news encoder architecture of some news recommenders.

Lastly, we contrast the representational similarity of models against their retrieval similarity. Fig. 5.3a illustrates the evolution of Jaccard similarity scores between the $\text{SE}_{\text{Monofeat}}$ encoder and the best performing architecture from each remaining news encoder family for different values of k. For



(a) $\text{SE}_{\text{Monofeat}}$ against the best performing architectures from the other news encoder families.

(b) LF against other user encoder architectures evaluated, with `CNN+AddAtt` as the base news encoder.

Figure 5.3: Evolution of Jaccard similarity for different values of k.

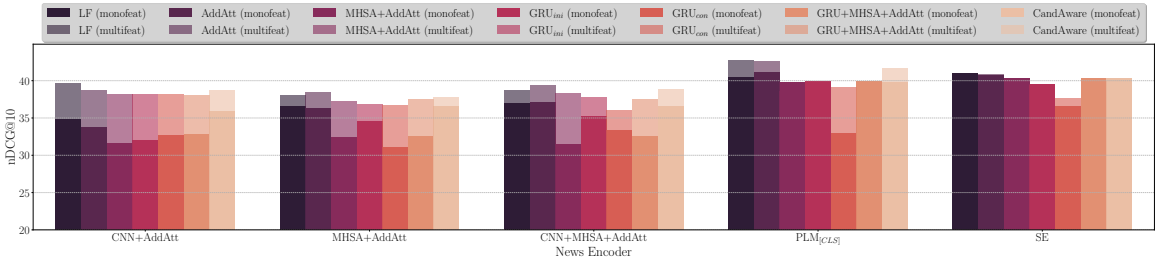
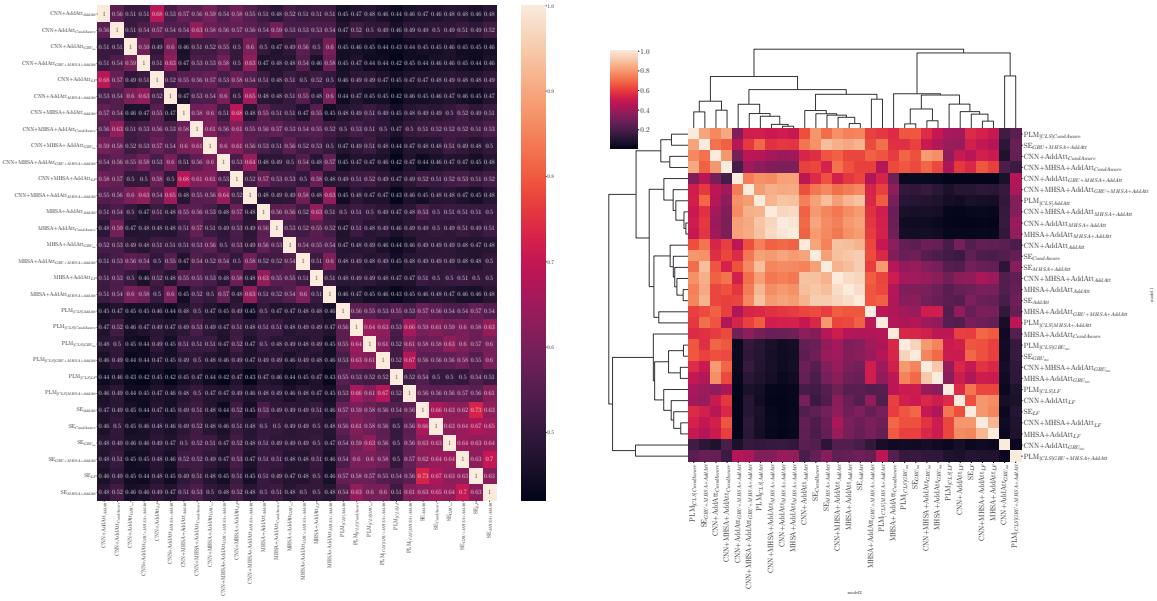


Figure 5.4: Ranking performance of different recommenders (nDCG@10) depending on the user encoder architecture, for different base news encoder families. The dark bars denote the ranking obtained when using a mono-feature input (i.e., title) in the news encoder, whereas the lighter bars indicate the (generally higher) scores gained with a multi-feature input (i.e., title and category), and the best multi-feature aggregation strategy per news encoder family.

small values of k, we observe a lower similarity of retrieved news for inter-family text encoders, with scores converging toward 1 for larger k. An important insight here is that for low values of k (e.g., $k < 10$), the news articles retrieved by different news encoders tend to be identical, on average, in more than half of the recommended items (e.g., a Jaccard of 0.42 for $k = 5$ translates into an overlap of 3 out of 5 items). We observe this behavior even for models with lower representational similarity scores, e.g., word embedding-based news encoders versus language model-based news encoders. This is relevant from a practical perspective, where retrieval similarity is of most interest for small values of k. It would imply, on the one hand, that the representational similarity of news encoders might not directly correlate with the retrieval performance for low k. On the other hand, this evidence re-affirms our earlier hypothesis that small differences in the architecture and complexity of news encoders do not result in large differences in the actual recommended items.

5.5.2 User Encoder Architectures

We next investigate the ranking performance, with regards to nDCG@10, for different user encoder architectures for the same base news encoder. Fig. 5.4 displays the corresponding results, for both mono-feature, and well as multi-feature input. We find that the `LF`, `AddAtt`, and `CandAware` encoders perform the best across all families of news encoders. More specifically, the much simpler `LF` and `AddAtt` encoders outperform the complex `CandAware` modeling technique in the case of language model-based news encoders, and perform similarly with `CandAware` for word embedding-based news encoders, as previously suggested by the results of Chapter 4. Surprisingly, these two approaches also consistently achieve better ranking than sequential-based user encoders (i.e., `GRU+MHSA+AddAtt`, GRU_{ini}, GRU_{con}). Once again, we see that using categorical information alongside the textual content



(a) Jaccard similarity of top-10 recommended news.

(b) CKA similarity of user embeddings.

Figure 5.5: Retrieval and representational similarity of models with different user encoder architectures. Each model name denotes the base news encoder, with the user encoder architecture indicated by the subscript.

as input to the news encoder benefits all recommenders regardless of the user encoder family. The only exception, as previously discussed, are SE-based NNRs. Interestingly, we see that multi-feature inputs close the gap (i) in between inter-family user encoders for the same base news encoder, and (ii) across intra-family user encoders for different underlying news encoder. Most importantly, our findings corroborate earlier results from Chapter 4 and Möller and Padó (2024) that the complexity of user encoders can be simplified, particularly when the bi-encoder NNR leverages language models pretrained, or even domain-specialized, on large-scale corpora, to obtain news representations.

The heatmap in Fig. 5.5a shows the Jaccard similarity scores for the top-10 recommendations, for the different user encoder families, when using only the title as input to the news encoder.[2] We exclude $\texttt{GRU}_{\text{con}}$ from further analysis as it underperforms the counterpart variant $\texttt{GRU}_{\text{ini}}$. We observe that in terms of retrieval similarity, the NNRs are clustered based on the underlying news encoder family, regardless of the user encoder used. Once again, the results indicate a large overlap of recommended news (i.e., on average, of at least 7 out of 10 recommendations) for the user encoders within these clusters. Moreover, we observe comparable similarity patterns across inter-family user encoders for the same news encoder family; different news encoders change only the absolute magnitude of the Jaccard similarity scores. Within intra-family clusters of news encoders, the findings re-affirm that `LF` and `AddAtt` have the highest overlap in terms of the top-10 recommended articles; their generated recommendations usually differ in at most 2 or 3 items, on average. This is intuitive, as `LF` represents a special case of `AddAtt`, where the attention weights are all equal, and set to the inverse of the user's history length.

We delve deeper into the retrieval similarity of user encoder architectures. Fig. 5.3b shows the Jaccard similarity of `LF` against the other user modeling approaches for a recommender with a `CNN+AddAtt`-based news encoder, for different values of k. As in § 5.5.1, the Jaccard similarity of recommended news is sensitive to the value of k, with scores converging toward 1 for larger values of k. On the one hand, the scores of sequential user encoders ($\texttt{GRU}_{\text{ini}}$, `GRU+MHSA+AddAtt`) are clustered closely together, which can be explained by their shared sequential component. However,

[2]The results with multi-feature input are similar, and we omit them for the sake of brevity.

the retrieved articles appear to be more similar between sequential and non-sequential user encoders (e.g., higher Jaccard similarity between GRU+MHSA+AddAtt and MHSA+AddAtt) across intra-family news encoders, than between sequential user encoders. This could be attributed to the architectural differences of the two models, among which GRU+MHSA+AddAtt employs an attention network similar to that of MHSA+AddAtt. These mixed results, combined with the better performing non-sequential user encoders, call into question the efficiency of modeling the news recommendation task as a sequential recommendation problem (Wu et al., 2022f).

We shift our attention to the pair-wise similarity of user embeddings generated by the different types of user encoders for the users in the test set, illustrated in the heatmap of Fig. 5.5b. We additionally perform a hierarchical clustering on the heatmap to identify clusters of similar user encoders (Waskom, 2021). In contrast to the retrieval results, we find that the architecturally comparable families of user encoders dictate the similarity of embeddings, regardless of the underlying news encoder used. Most surprisingly, we find that although the top-recommended news by GRU_{ini} and GRU+MHSA+AddAtt moderately overlap, their user representations are highly dissimilar. Moreover, the latent representations of AddAtt appear more similar to other attention-based user encoders than with LF. This could be explained by the fact that as a particular case of AddAtt, the parameterless LF does not reshape the embedding space, as it simply computes an average of the user's clicked news. Nonetheless, these differences in the representational similarities of user encoders also do not appear to directly correlate with more dissimilar retrieval performance. This suggests that in real-world applications, the lightweight and conceptually simple LF constitutes an equally effective and more efficient alternative to AddAtt, and especially, to more complex architectures.

5.5.3 Key Takeaways

Following the results of our in-depth analysis of the embedding and retrieval similarity of the most prominent news and user encoder architectures, we highlight several key takeaways.

Semantic Quality of News Representations is Key. Our analysis demonstrates that the semantic accuracy of representations produced by the news encoders, achieved either through multi-feature input or contextualized language models, significantly outweighs the impact of user encoders. This is particularly the case when initializing news representations with large-scale PLMs. Additionally, contextualized language models can effectively capture semantic nuances, such as topical information, without heavily relying on categorical annotations. From a practical standpoint, this reduces the need for manual or automatic feature engineering, streamlining the NNR design process. We hence argue that research on news encoding should focus more on leveraging and adapting existing semantically informed, contextualized language models for the task of news recommendation, rather than on incrementally modifying existing architectures.

User Encoders Can be Considerably Simplified. Our findings show that retrieval similarity is primarily influenced by the underlying news encoder family, rather than the specific user encoder used. At the same time, simpler approaches, such as LF and AddAtt, not only result in significantly better ranked results, but their retrieved items largely overlap with those recommended by more complex user encoder architectures. These findings thus render simpler architectures as better and more lightweight user modeling alternatives. Additionally, the high retrieval similarity between parameter-free (i.e., LF) and parameterized (e.g., AddAtt) encoders heavily indicates that, in practice, there is little empirical justification for an additional parameterized component in the news recommender system. Furthermore, the similarity of sequential and non-sequential encoders indicates that treating news recommendation as a sequential problem might be sub-optimal. We speculate that the high item churn characteristic of news, combined with short user histories, limit the benefits of differentiating between long and short-term user preferences, in contrast to other domains, such as movie or book recommendation. In conclusion, in line with Möller and Padó (2024), we posit that user modeling should not focus exclusively on the architectural component, but instead, should pay closer attention

to the users' motivations to consume certain news, on the one hand, and to collecting richer and more accurate user (relevance) feedback, on the other hand.

More Rigorous Evaluation is Needed for Better Model Selection. Our findings, along with recent research (Möller and Padó, 2022, 2024, Chapter 4), highlight the limitations of current evaluation practices in news recommendation. By focusing solely on performance metrics, we risk overlooking critical aspects of model behavior, leading to sub-optimal component selection and incremental model advancement. Therefore, we advocate for a more comprehensive and rigorous evaluation approach. Ablation studies should consider the broader architectural context, and together with model comparisons, should extend beyond performance-based evaluation to include a more granular behavioral and representational analysis. This would provide a more nuanced understanding of model similarities and differences, guiding researchers and practitioners toward better informed model selection decisions.

5.6 Conclusion

Despite the central role played by encoder architectures in neural news recommenders, their advancement and understanding is generally limited to one-sided evaluation in terms of recommendation performance. In this Chapter, we conducted a comprehensive evaluation of encoder architectures in neural news recommenders, by systematically analyzing their (i) representation similarity, (ii) overlap of generated recommendations, and (iii) overall recommendation performance. Evaluations of recommenders on standard benchmarks often reveal insignificant performance differences between compared models or among their ablated components. Consequently, our analysis of differences in representational similarity and retrieval overlap of neural news recommenders serves as a complementary evaluation tool for understanding the relationship between the architectural design, behavior, and downstream performance of models.

Our findings offer more nuanced insights into the interplay of news and user encoders, and challenge the assumption that complex encoding techniques are essential for accurate news recommendation. We demonstrate that simpler, yet equally effective architectures can yield comparable results. This underscores the importance of understanding recommenders' behavior from multiple perspectives, and of balancing model complexity with performance. Specifically, we emphasize three key takeaways: (1) the crucial role of the semantic accuracy of news representations, (2) the potential for simplifying user encoders without sacrificing accuracy, and (3) the need for more rigorous evaluation and ablation studies to inform architectural design choices. By fostering a more transparent and nuanced understanding of encoder architectures in neural news recommenders, we hope to guide researchers and practitioners toward more efficient and effective model designs.

In this Chapter and the previous one, we have conducted a comprehensive and multidimensional analysis of the common components of neural news recommendation models, encompassing user modeling paradigms, training objectives, and the encoder architectures underlying both the news and user representations (**C1**). Across both Chapters, we demonstrated that the neural architectures of NNRs can be substantially simplified without sacrificing performance, and we derived several key takeaways for designing more effective and efficient news recommenders. In the next Chapter, we turn our attention to investigating the effects of bias in algorithmic news curation (**C2**).

Part III

Multi-aspect Bias Analysis and Mitigation

Chapter 6

Multi-Aspect Bias Analysis in News Recommenders

Every way of seeing is also a way of not seeing.

Kenneth Burke
«Permanence and Change »

* News recommendation plays a critical role in shaping the public's worldview through the way in which it filters and disseminates information across a wide range of topics, as discussed in § 1.1.2. Algorithmic news curation has been hypothesized to contribute to the formation of filter bubbles and to intensify users' selective exposure, potentially increasing their susceptibility to polarized opinions and misinformation. Given the crucial impact of media in opinion formation, particularly in politically or socially sensitive topics, understanding the societal effects of personalized news recommendation beyond accuracy has become essential in today's digital society. Recommender systems that are not designed responsibly may inadvertently amplify the voices of certain actors while further marginalizing under-represented communities.

Motivated by these concerns, this Chapter presents a comprehensive analysis of the effects of algorithmic news curation on users' selective exposure (**C2**). We proceed in two stages. First, we investigate whether relatively simple news recommender systems exhibit biases toward content with specific emotional valence (sentiment) or perspective (stance) when dealing with a sensitive topic. To support this analysis, we compile GeNeG, a German news corpus on the topic of migration, comprising articles from media outlets spanning the full political spectrum and annotated with sentiment polarity and stance labels. Using GeNeG, we show that recommender systems tend to favor content aligned with particular sentiment and stance patterns, reflecting both properties of the articles themselves and users' prior reading behavior.

Building on these findings, we extend GeNeG into NeMig, a bilingual German-English news collection by additionally gathering English-language articles from US publishers covering the same topic and representing diverse political orientations. Beyond sentiment, NeMig contains annotations for media outlet political leaning, subtopics, and disambiguated named entities. These enriched annotations enable analyses of algorithmic news curation that go beyond accuracy-focused evaluation, including the study of recommender bias, selective exposure, and filter-bubble effects. Crucially,

This Chapter is adapted from: (1) Mehwish Alam, **Andreea Iana***, Alexander Grote, Katharina Ludwig, Philipp Müller, and Heiko Paulheim. 2022. Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 448–457, New York, NY, USA. Association for Computing Machinery and (2) **Andreea Iana**, Mehwish Alam, Alexander Grote, Nevena Nikolajevic, Katharina Ludwig, Philipp Müller, Christof Weinhardt, and Heiko Paulheim. 2023a. NeMig - A Bilingual News Collection and Knowledge Graph about Migration. In *INRA 2023: Proceedings of the International Workshop on News Recommendation and Analytics co-located with the 2023 ACM Conference on Recommender Systems (RecSys 2023)*, volume 3561, pages 1–22, Aachen, Germany. RWTH Aachen.

unlike most existing datasets that rely solely on implicit click feedback, NeMig includes anonymized user socio-demographic and political information. Using NeMig, we extend our bias analysis to widely used neural news recommenders, focusing on three central aspects of news content: sentiment, topic, and political orientation. Our results confirm that NNRs also exhibit systematic biases with respect to content characteristics and users' interaction histories. At the same time, we find that diversity-aware and knowledge-aware modeling approaches can help mitigate certain forms of aspect-based bias, albeit sometimes at the cost of predictive performance.

Overall, GeNeG and NeMig provide valuable resources for the news recommendation community and aim to foster interdisciplinary research into the multidimensional societal effects of algorithmic news curation.

6.1 Introduction

Every day, large volumes of news articles are being published online, leading to an information overload that exceeds the consumptive capacities of Internet users. To alleviate this and deliver personalized content, online news platforms increasingly rely on recommender systems that tailor suggestions to users' past interests and behavior. However, optimizing recommendations for user preference alignment often results in highly similar content being repeatedly suggested (Wu et al., 2020c; Liu et al., 2021b). Arguably, such algorithmic news curation may limit users' exposure to diverse content by filtering out articles deemed irrelevant (in order to maximize user engagement) (Pariser, 2011) or inconsistent with preexisting beliefs and attitudes, as users tend to favor information reinforcing their opinions (Freedman and Sears, 1965).

Over time, the power of recommender systems to shape users' perception of the world has led to concerns that individuals are being isolated from diverse perspectives through algorithmically created "filter bubbles" (Pariser, 2011) and self-selected "echo chambers" (Sunstein, 2001; Jamieson and Cappella, 2008), in which individuals primarily encounter homogeneous ideological viewpoints. In politically or socially sensitive contexts (e.g., war, climate change, migration), an over-exposure to less diverse perspective may foster an attitude-reinforcing spiral (Stroud, 2017; Donsbach and Mothes, 2013) which, in the long run, can contribute to polarization, or even radicalization of individuals with extreme political or ideological viewpoints (Bakshy et al., 2015; Boutyline and Willer, 2017; Liu et al., 2021b, *inter alia*).

This influence of recommender systems on news exposure underscores the importance of *diversity*, a cornerstone of both media quality (McQuail, 1992) and democratic discourse, to ensure a balanced and well-informed public sphere in any democratic society (Helberger, 2021). News diversity is nonetheless a multi-dimensional concept (Haim et al., 2018), encompassing the pluralism of sources (*source diversity*) (Voakes et al., 1996), of topics (*content diversity*) (Haim et al., 2018), and of stances on a given topic (*viewpoint diversity*) (Baden and Springer, 2017). While past research has extensively addressed personalization and recommendation accuracy (Wu et al., 2023), resources for media bias detection and analysis (Spinde et al., 2020; Lim et al., 2020; Färber et al., 2020), or interdisciplinary connections between media and computer science (Hamborg et al., 2019), few studies have examined how news recommenders affect exposure to such multidimensional diversity or how algorithmic biases manifest in the recommended content.

From the perspective of recommender-system analysis, *sentiment* and *stance* constitute two valuable dimensions for quantifying and mitigating bias algorithmic bias. Sentiment analysis identifies whether a piece of text expresses a positive, neutral, or negative attitude (Liu et al., 2010a; Ravi and Ravi, 2015), while stance detection identifies an article's or author's position toward a given target issue (Küçük and Can, 2020; Reuver et al., 2021). Sentiment information is leveraged by fairness-aware recommenders to produce sentiment-diverse or sentiment-agnostic suggestions (Wu et al., 2020c, 2022h). Likewise, sentiment enables the investigation of whether algorithmic news curation propagates or amplifies readers' polarization (McCoy et al., 2018; Pariser, 2011; Cho et al.,

2020), or reflects systematic biases present in media ecosystems (Gangula et al., 2019; Lazaridou and Krestel, 2016). Stance and sentiment annotations can therefore serve as proxies for viewpoint diversity and indicators of polarization. Moreover, these annotations can be leveraged to design news recommenders that explicitly promote diversity across ideological or affective dimensions.

To explore these issues, we conduct a series of preliminary analyses on news recommenders, investigating whether different algorithms exhibit underlying biases towards particular sentiments or viewpoints, thereby reducing diversity and intensifying selective exposure (Pitoura et al., 2021). To this end, we construct GeNeG, a dataset of German news articles on migration, annotated via transfer learning with sentiment scores and stance labels. Using this dataset, we examine whether recommendation outputs exhibit systematic biases aligned with article sentiment, stance, or users' reading histories. These analyses shed light on the ways in which algorithmic personalization may intensify selective exposure and reduce the diversity of recommended content.

Despite advances in bias quantification, progress remains constrained by the lack of suitable datasets for analyzing recommender effects beyond accuracy performance and across languages. Existing news datasets for recommendation (i) largely focus on general or less sensitive topics, such as entertainment, fashion, or sports (Gulla et al., 2017; Wu et al., 2020e) and (ii) omit features critical for studying "filter bubbles" (Pariser, 2011) or algorithmic bias (e.g., emotional valence, political orientation). Conversely, datasets for media analysis, such as those for fake news detection (Wang, 2017; Horne et al., 2018), news narratives analysis (Shu et al., 2020; Levi et al., 2020), or news bias detection (Färber et al., 2020; Lim et al., 2020), typically lack user information, preventing their use in recommendation scenarios.

To address these gaps, we expand the GeNeG dataset and introduce NeMig, a bilingual news collection and set of knowledge graphs (KGs) on refugees and migration, in German and English, accompanied by real-world user data. The corpus contains over 7,000 German and 10,000 US news articles spanning a wide political spectrum. **1)** Unlike existing resources, NeMig (i) targets the same polarizing topic in two languages and (ii) is annotated for sentiment polarity, media outlet political leanings, and subtopics. **2)** Each article is enriched with disambiguated named entities linked to Wikidata, which we leverage to construct corresponding KGs that integrate textual, metadata, and relational information, further expanded with up to two-hop neighbors from Wikidata of the named entities. **3)** NeMig also provides anonymized explicit feedback and socio-demographic data for approximately 3,000 users per language. Although smaller than benchmarks such as MIND (Wu et al., 2020e) or EB-NeRD (Kruse et al., 2024), NeMig's inclusion of users' media consumption, political attitudes and interests, personality traits and demographics, enables research on recommender systems' effects beyond accuracy-based performance, e.g., on political polarization. Furthermore, its user data supports the generation on synthetic user datasets, comprising not only click behavior information, but also explicit data about the users' background and preferences.

We demonstrate the utility of NeMig through a series of experiments. First, we benchmark several news recommenders on the dataset, evaluating their performance in terms of accuracy-based metrics. Second, we investigate biases and polarization effects in model outputs using diversity-related metrics. Lastly, we explore the contribution of various features to the quality of the corresponding KGs.

6.2 Related Work

News Recommendation Benchmarks. Numerous datasets have been developed for trainin and benchmarking news recommender systems, including Plista (Kille et al., 2013), Adressa (Gulla et al., 2017), Globo (de Koning et al., 2018; Gabriel De Souza et al., 2019), MIND (Wu et al., 2020e), EB-NeRD (Kruse et al., 2024)). These large-scale monolingual benchmarks provide extensive user interaction data, bot typically lack features essential for analyzing recommender effects beyond accuracy, such as news-level attributes (e.g., sentiment orientation, political leaning) and user-level information (e.g., demographics, political interests). In contrast, GeNeG and NeMig provide (i)

monolingual, and respectively, bilingual news collections, annotated not only with text and metadata information, but also with subtopic, sentiment and political orientation, and (ii) implicit user feedback, enriched with demographic and political data.

News Knowledge Graphs. Knowledge graphs have proven effective in modeling semantic relations between news articles that purely text-based methods fail to capture (Wang et al., 2018b; Iana et al., 2024). Within the Newsreader[1] project, focused on multilingual processing of news articles, event-centric KGs were generated to describe what events occurred, where an when they took place, and who was involved (Vossen et al., 2016; Rospocher et al., 2016). Similarly, News Graph (Liu et al., 2019a) was constructed specifically for news recommendation, using a large MSN News corpus comprising over 621K articles, 594K news entities, and user-item interaction logs. The graph integrates news content, user behaviors, and topic entities enriched with neighboring triples from Microsoft Satori, and includes three types of collaborative entity relations: co-occurrence in the same news, co-clicks by the same user, and co-clicks by the same user within the same browsing session. A comprehensive overview of how semantic KGs support news production, distribution, and consumption is provided by Opdahl et al. (2022). Unlike News Graph, which depends on a proprietary knowledge base, GeNeG and NeMig employ the open-source Wikidata for entity linking, fostering transparency and reproducibility.

News Datasets for Media Discourse Analysis. The ubiquity of online news has made it a central resource for media discourse analysis. For instance, Horne et al. (2018) introduced a large dataset of political news from mainstream and alternative outlets, enriched with content-based and social media engagement features, suitable for studying news and engagement characterization, news attribution, content copying, or exploration of news narratives. Similarly, Lim et al. (2020) and Färber et al. (2020) collected and annotated articles with different types of bias fr sentence-level bias analysis. Other efforts focus on fake news detection, including monolingual (Wang, 2017; Shu et al., 2020) and multilingual cross-domain datasets (Shahi and Nandini, 2020), or differentiating fake from satirical news content (Golbeck et al., 2018; Liu et al., 2019c). Although these news collections provide rich annotations (e.g., bias, social media features, social context, spatio-temporal information), they are ill-equipped to be seamlessly utilized in training recommendation algorithms due to the lack of user feedback data. Both GeNeG and NeMig address this gap by combining sentiment and political leaning annotations with comprehensive user feedback and socio-demographic information.

6.3 GeNeG: German News Dataset

6.3.1 News Data Collection and Annotation

We construct the news knowledge graph GeNeG (Iana et al., 2022b) from a dataset of 4,557 German news articles on the topic of refugees and migration, collected from 39 media outlets spanning a broad political spectrum, including both far-right and far-left online publications. The corpus comprises articles published between January 1st, 2019 and October 20th, 2020, selected using keyword stems representative of the topic (e.g., 'flüchtl', 'migrant', 'asyl'), and gathered via a custom web crawler.[2] The dataset includes rich textual and metadata information, such as article content, provenance, publication date, authorship, or tagged keywords. Named entities were extracted from both article content (e.g. persons, locations) and metadata (e.g. publisher, author) and subsequently disambiguated using Wikidata (Vrandečić and Krötzsch, 2014).

The resulting GeNeG news knowledge graph is a heterogeneous network in which news articles and real-world entities are represented as nodes, while different relations between them form the edges. Nodes are divided into two categories: *literals*, denoting textual content, dates or polarization scores, and *entities*, indicating identified named entities. Entity nodes may be either linked to Wikidata

[1] http://www.newsreader-project.eu

[2] The implementation of the custom web crawlers is available at https://github.com/andreeaiana/german-news.

Knowledge Graph Variant	Nodes	Edges	Properties
base GeNeG	54,327	186,584	16
entities GeNeG	844,935	6,615,972	1,263
complete GeNeG	868,159	6,656,779	1,271

Table 6.1: GeNeG knowledge graph statistics.

	Stance Class		
Subset	Against	Favor	Total
Training	17130	16720	33850
Validation	1451	1420	2871
Test	5882	6009	11891

Table 6.2: Statistics of the the German subset of x-stance.

entries or remain non-disambiguated when no corresponding Wikidata item is found. GeNeG is released in three variants: (i) *base graph* – containing textual information, metadata, and extracted entities; (ii) *entities graph* – derived from the base version by removing all literal nodes, and enriching the remaining entity nodes with up to three-hop neighbors from Wikidata' (iii) *complete graph* – combining the previous two variant, integrating both literals and entities. Table 6.1 summarizes the statistics of these three KG variants in terms of the number of nodes, edges, and properties.

Sentiment Annotations. We use a pretrained BERT-based German-language sentiment classification model (Devlin et al., 2019; Guhr et al., 2020) to annotate each article with a sentiment score. The model applies a softmax layer to produce probability estimates for three classes: positive, neutral, and negative. We derive a continuous sentiment score in the range $[-1, 1]$ by computing the difference between the positive and negative class probabilities, defined as $s = p_p - p_n$, where p_p and p_n denote the probabilities of the positive and negative sentiment, respectively. The neutrality score is omitted, as it is implicitly captured by $1 - p_p - p_n$. On the GeNeG corpus, the model reveals a slight skew toward negative sentiment, with a mean score of $s_{mean} = -0.154$ and a median of $s_{median} = -0.019$.

Stance Annotations. We employ a transfer learning approach to perform stance detection on the news articles in our corpus. The task involves classifying each article into one of two categories: 'in favor' or 'against'. Among the available German-language stance detection resources (Mascarell et al., 2021; Vamvas and Sennrich, 2020), we selected the x-stance dataset for model training. x-stance (Vamvas and Sennrich, 2020) facilitates multilingual, multi-target stance detection using a Multilingual BERT (Devlin et al., 2019) model and contains question-article pairs in English, French, German, and Italian. The dataset provides a balanced distribution of instances across the 'in favor' and 'against' classes, which makes it suitable for transfer learning. For our experiments, we extracted only the German subset (language label: "de"). The statistics of this subset are shown in Table 6.2.

Stance detection is formulated as a classification task over question-article pairs, where the stance of the article is predicted relative to a predefined question. The questions used in our setup are listed in Table 6.3. We fine-tuned GermanBERT[3] (Chan et al., 2020) on the German portion of x-stance using a learning rate of $3e-5$ and four training epochs. We then classify the GeNeG articles using the fine-tuned model into the two stance categories ('in favor' or 'against'), according to the predefined questions. Table 6.4 reports the stance detection results for each question, as well as the overall average score, computed as $\frac{favor-against}{favor+against}$.

[3] https://huggingface.co/bert-base-german-cased

ID	German Question	English Translation
Q1	Befürworten Sie, dass Flüchtlinge nach Deutschland kommen?	Are you in favor of refugees coming to Germany?
Q2	Befürworten Sie, dass Flüchtlinge in Deutschland leben?	Are you in favor of refugees living in Germany?
Q3	Befürworten Sie, dass Flüchtlinge in Deutschland arbeiten?	Are you in favor of refugees working in Germany?
Q4	Sollte Deutschland Flüchtlinge aufnehmen?	Should Germany take in refugees?
Q5	Sollte Deutschland Flüchtlingen helfen?	Should Germany help refugees?

Table 6.3: Questions for the question-news article pairs for the stance detection task.

ID	# Articles in Favor	# Articles Against	Avg. Score
Q1	2165	2392	-0.050
Q2	2193	2364	-0.038
Q3	2210	2347	-0.030
Q4	2120	2437	-0.070
Q5	2192	2365	-0.038

Table 6.4: Number of articles per stance and pairing question.

Finally, we enrich the GeNeG knowledge graph with the stance labels by introducing two new properties: `geneg:in_favor` and `geneg:against`. Each stance annotation is represented as a triple of the form $\langle article, stance, question \rangle$, where $stance \in \{\texttt{geneg:in_favor}, \texttt{geneg:against.}\}$ and $question \in \{Q_n, n = \overline{1,5}\}$. The resulting triples link each article to its predicted stance relative to a specific question, thereby extending the graph with interpretable stance-level relations.

Data Availability. GeNeG is available on Zenodo under a CC BY-NC-SA 4.0 license.[4] Access to the dataset is granted upon request, in accordance with copyright restrictions on the full news articles.

6.3.2 User Data Collection

Recommendation Models. We use four content-based recommender systems, for two purposes: first, to collect user interaction data during an online study, and second, to conduct subsequent offline experiments. The first thee recommenders are *text-based*, relying solely on article content, while the fourth is *knowledge-based*, leveraging the graph representations of the GeNeG knowledge graph. In all text-based models, we compute the relevance of an article to a user's reading history using cosine similarity. **(1)** The **TF-IDF** recommender represents each article as a vector of term frequency-inverse document frequency values (Schütze et al., 2008). **(2) word2vec** encodes articles using distributed word representations obtained from a pretrained word2vec model trained on Common Crawl and Wikipedia (Mikolov et al., 2013; Grave et al., 2018), with the article's embedding computed as the average of its word embeddings. **(3)** The **Transformer** model leverages contextual embeddings from a pretrained cross-lingual sentence transformer (Reimers and Gurevych, 2019; May, 2020), where the representation of each article is obtained by averaging the embeddings of its sentences. **(4) RippleNet** (Wang et al., 2018a) is a knowledge-aware recommender that propagates user preferences along the edges of a knowledge graph through *ripple sets*. These sets capture higher-order connections (i.e., multi-hop entity neighbors) between entities extracted from the user's previously read articles,

[4] https://doi.org/10.5281/zenodo.6039372

Dataset	Training	Complete Test	Random Test	Total
Items	3,365	1,633	316	3,825
Users	1,414	1,174	177	1,417

Table 6.5: GeNeG user data statistics.

enabling the model to infer the likelihood of interest in new content. The strength of the user's preference diminishes proportional to the distance from the original seed in the knowledge graph. The final click probability is predicted using the preference distribution of the user for a candidate news, obtained by superposing multiple ripple sets. We train RippleNet on the *entities GeNeG* graph, excluding sentiment and stance information.

User Data. We collect user data through an online study designed to assess the potential polarization effects of news recommender systems on users. The study uses the GeNeG corpus (§ 6.3.1), from which we exclude 732 articles longer than 1,500 words to ensure manageable reading times and maintain participant engagement.

We recruit participants via the Respondi AG online panel using quota sampling to reflect the demographic structure of German Internet users aged 18 to 74. The final dataset comprises 1,417 valid participants (from 1,801 initially recruited), with a mean age of 46.56 years (SD=15.68), 47.7% female, and 38.5% holding an *Abitur* (German university entrance certificate) or higher qualification. Only desktop users were included to ensure consistent display conditions. We further excluded participants who did not meet the age quota, spent less than 120 seconds or more than 2.5 hours on the platform, or exhibited uniform response patterns across all survey items. Participants first completed a questionnaire measuring political orientation and interests, attitudes towards refugees, and demographics. Afterwards, each participant was randomly assigned to one of the following four recommenders: TF-IDF, word2vec, Transformer, or a random recommendation baseline, and completed four interaction rounds. In each round, participants selected one article from a preview of six and then read it, resulting in four logged interactions per user. At the end of the experiment, a post-survey was administered to assess potential changes in political polarization. For further methodological details, we refer the reader to Ludwig et al. (2023).

We split the resulting dataset into a training and a test portion (denoted *complete test*), using an 80:20 ratio. Because 85% of the user ratings originate from text-based recommenders, we mitigate the potential overfitting bias of these models by constructing an additional test subset (denoted *random test*) containing only user ratings for randomly recommended articles. Table 6.5 presents descriptive statistics of the user data.

6.4 Preliminary Bias Analyses

6.4.1 Experimental Setup

Benchmarked Recommenders. We evaluate the four recommendation models described in § 6.3.2 – TF-IDF, word2vec, Transformer, RippleNet– on GeNeG. We report performance using Accuracy (ACC), AUC, and the F_1-score.

Bias Analysis. Beyond predictive accuracy, we examine whether the recommenders exhibit stance of sentiment bias. Each text-based model returns the top k most similar articles to those in a user's reading history, while RippleNet recommends the k articles with the highest predicted click probability. For comparability, in our bias analysis experiments, we fix $k = 5$, reflecting the common practice of news outlets to suggest approximately five related articles per page.

To quantify article stance and compute bias scores for users and recommenders, we map stance

	Test Set					
	Complete			Random		
Model	ACC	AUC	F_1	ACC	AUC	F_1
TF-IDF	**0.73**	**0.87**	0.65	0.49	0.50	0
Word2vec	0.51	0.79	**0.67**	0.50	0.47	0.66
Transformer	0.51	0.78	0.67	0.50	0.52	**0.67**
RippleNet	0.55	0.57	0.52	**0.56**	**0.58**	0.53

Table 6.6: Click-through rate (CTR) prediction results.

labels to numerical values as follows:

$$stance_score = \begin{cases} +1 & \text{if } stance_label = \text{Favor} \\ -1 & \text{if } stance_label = \text{Against} \end{cases} \tag{6.1}$$

For each user, we calculate an average sentiment bias score by taking the mean sentiment score of the articles in their reading history. Similarly, for each user-recommender pair, we compute the recommender's sentiment bias score as the mean sentiment score of the recommended articles; averaging across users yields the recommender's overall sentiment bias. Stance bias scores are computed analogously. All bias scores lie in $[-1, 1]$, where -1 indicates a tendency toward negative sentiment or stances against a topic, and $+1$ indicates the opposite.

Implementation and Optimization Details. All recommenders are optimized for click-trough-rate (CTR) prediction.[5] Each recommender estimates, for every user-article pair in GeNeG test set, the likelihood that the user will click on the candidate article. For all text-based recommenders, we use min-max scaling to approximate probability values from the models' similarity score outputted by the model. For the TF-IDF model, we use an n-gram range of 1 and 2, together with L2-norm regularization. The word2vec and the Transformer models represent articles using 300-dimensional and 768-dimensional embeddings, respectively. For RippleNet, we set the number of propagation hops to $H = 1$, the ripple set size to 16, the item and knowledge graph embedding dimensionality to $d = 48$, and the training weight for knowledge graph embeddings to $\lambda_2 = 0.03$. All remaining hyperparameters follow the default news-recommendation settings from Wang et al. (2018a).

6.4.2 Results and Discussion

CTR Results. Table 6.6 reports the accuracy-based performance of the four recommenders. On the *complete test* set, the text-based recommenders achieve the highest AUC and F_1-score, while RippleNet outperforms the word2vec and Transformer models in terms of accuracy. However, this evaluation setting is biased in favor of text-based models: over 55% of all user interactions originate from participants assigned to TF-IDF conditions, leading to overfitting on text-driven preferences.

To mitigate this bias, we further evaluate all models on the *random test* set, which contains only interactions where article suggestions were sampled uniformly at random. In this setting, RippleNet outperforms all other recommenders in both accuracy and AUC, and achieves a higher F_1-score than TF-IDF, which fails to correctly predict any positive examples and therefore, yields an F_1-score of 0. Overall, these results indicate that RippleNet, relying only on entity-level information from GeNeG, achieves competitive performance and demonstrates greater robustness to variations in interaction data than purely text-based recommenders.

Recommender-Data Bias Analysis. We first examine the average recommender bias score to address

[5]Code and data for the experiments are available at https://github.com/andreeaiana/geneg_benchmarking.

Test set	Avg. user sentiment score	Avg. recommender sentiment score			
		TF-IDF	Word2vec	Transformer	RippleNet
Complete	-0.171 (*)	-0.162 (-/-)	-0.169 (-/**)	-0.157 (-/-)	-0.148 (*/-)
Random	-0.169 (-)	-0.141 (-/-)	-0.170 (-/-)	-0.160 (-/-)	-0.150 (**/-)

Table 6.7: Average sentiment scores for users and recommenders. Statistical significance (Student's t-test) is marked as * ($p - value < 0.01$) and ** ($p - value < 0.05$). For user sentiment, a star indicates significance relative to the corpus average. For recommender sentiment, $(/)$ indicates significance relative to the user and corpus averages, respectively. A "–" denotes no significant difference.

ID	Test Set	Avg. user stance score	Avg. recommender stance score			
			TF-IDF	Word2vec	Transformer	RippleNet
Q1	Complete	-0.109 (-)	-0.140 (-/-)	-0.165 (**/-)	-0.136 (-/-)	-0.082 (-/-)
	Random	-0.093 (-)	-0.227 (**/–)	-0.219 (**/-)	-0.172 (-/-)	-0.054 (-/-)
Q2	Complete	-0.102 (**)	-0.132 (-/-)	-0.158 (**/-)	-0.131 (-/-)	-0.074 (-/-)
	Random	-0.093 (-)	-0.220 (**/–)	-0.207 (**/-)	-0.169 (-/-)	-0.038 (-/-)
Q3	Complete	-0.092 (**)	-0.127 (-/**)	-0.149 (**/**)	-0.116 (-/**)	-0.062 (-/**)
	Random	-0.081 (-)	-0.215 (**/–)	-0.205 (**/-)	-0.164 (-/-)	-0.024 (-/-)
Q4	Complete	-0.117 (-)	-0.167 (**/-)	-0.178 (**/-)	-0.157 (-/-)	-0.095 (-/-)
	Random	-0.106 (-)	-0.255 (**/–)	-0.268 (*/-)	-0.179 (-/-)	-0.084 (-/-)
Q5	Complete	-0.079 (-)	-0.130 (**/-)	-0.135 (**/-)	-0.124 (-/-)	-0.060 (-/-)
	Random	-0.081 (-)	-0.237 (*/–)	-0.199 (**/-)	-0.143 (-/-)	-0.055 (-/-)

Table 6.8: Average stance scores for users and recommenders. Statistical significance (Student's t-test) is marked as * ($p - value < 0.01$) and ** ($p - value < 0.05$). For user stance, a star indicates significance relative to the corpus average. For recommender stance, $(/)$ indicates significance relative to the user and corpus averages, respectively. A "–" denotes no significant difference.

the question: *do recommender systems tend to suggest articles with a particular sentiment or stance?*

Table 6.7 shows the results for *sentiment bias*. All models show a tendency to recommend articles with negative sentiment. This skew aligns with the overall sentiment distribution in GeNeG, where articles are, on average, negatively toned (cf. § 6.3.1). However, the text-based recommenders exhibit a stronger inclination toward negatively framed news: in five out of six settings, their average sentiment bias is more negative than the corpus average ($s_{mean} = -0.154$). In contrast, the knowledge-aware model consistently recommends articles with slightly less negative sentiment than the dataset mean. We observe the only statistically significant result ($p - value < 0.05$) for the word2vec model on the *complete test* set.

Table Table 6.8 illustrates the *stance bias* findings across both test sets. All recommenders favor articles expressing stances against migration and refugees across all questions. This mirrors the stance imbalance in GeNeG (cf. Table 6.4), where articles opposing the topic are more frequent. Still, the recommenders generally amplify this skew, producing recommendation lists with more negative stance bias averages than the corpus mean. As for sentiment analysis, RippleNet shows the weakest negative bias, deviating only sightly from the overall corpus-level stance scores. Nonetheless, we find statistical significance only for question Q3.

Recommender-User Bias Analysis. We next investigate: *to what extent do the recommenders'*

Bias Case	Test Set	TF-IDF	Word2vec	Transformer	RippleNet
C1	Complete	1166	1166	1165	1160
	Random	177	177	177	177
C2	Complete	8	8	9	14
	Random	0	0	0	0

Table 6.9: Recommender-user sentiment correlation counts.

Test Set	TF-IDF	Word2vec	Transformer	RippleNet
Complete	0.325 (*)	0.347 (*)	0.216 (*)	0.045 (**)
Random	0.253 (*)	0.310 (*)	0.129 (-)	-0.033 (-)

Table 6.10: Recommender-user sentiment bias Pearson correlation. Statistical significance is marked with * ($p - value < 0.01$) and ** ($p - value < 0.05$). A '-' denotes no significant difference.

sentiment and stance biases correlate with the existing user bias?. This analysis examine whether recommenders *align* or *amplify* users' preferences. Tables 6.7 and 6.8 report the users' average sentiment and stance bias scores.

Sentiment Bias Correlation. Users exhibit a slight preference for negatively toned articles (cf. Table 6.7), beyond what can be explained by the dataset's overall negative sentiment skew ($s_{mean} = -0.154$). The average user sentiment bias is -0.171 on the *complete test* set, and -0.169 on the *random test* set, both statistically significantly lower than the corpus mean, according to the Student's t-test. In seven out of eight cases, recommenders suggest articles with sentiment scores slightly less negative than the average user sentiment bias, although the difference is small. This suggests that recommenders largely reflect users' sentiment tendencies. We find the only statistically significant difference in the case of RippleNet, on the *complete test* set, indicating that the knowledge-aware model may be less prone to reinforcing existing sentiment preferences. To further analyze these bias patterns, we define two cases of user-recommender sentiment bias correlation:

- *(C1)* Same-direction bias: both the user and the recommender have a tendency toward articles with negative (positive) sentiments.
- *(C2)* Opposing-direction bias: the user has a preference for articles with negative (positive) sentiment, while the recommender tends to suggest news with the opposite sentiment.

We note that neutral cases are not possible given the nature of sentiment scores and the lack of perfectly neutral articles in the corpus. Table 6.9 shows the number of users classified in each of these cases. We find that almost all users fall into *(C1)*, meaning that recommenders overwhelmingly reinforce users' preference with regards to the tone of articles.

Pearson correlations, shown in Table 6.10, confirm these findings. TF-IDF and word2vec show significant positive recommender-user sentiment bias correlation on both test sets' the Transformer recommender does so only on the *complete test* set. That means that, if a user prefers articles with a negative tone towards the topic of refugees living in Germany, these recommenders will continue suggesting news with a similarly negative sentiment, thus reducing the user's exposure to articles with diverse sentiments. In contrast, RippleNet shows significance only on the *complete test* set ($p - value < 0.05$) and no significant on the *random test* set, suggesting that this model is less prone to amplifying users' existing sentiment preferences.

Stance Bias Correlation. A similar pattern emerges for stance. Text-based recommenders generally exacerbate users' preferences for articles against refugees and migration, with this amplification stronger on the *random test* set. word2vec exhibits significant differences compared to the average

		Complete Test Set				Random Test Set			
ID	Bias Case	TF-IDF	Word2vec	Transformer	RippleNet	TF-IDF	Word2vec	Transformer	RippleNet
Q1	C1	600	575	575	470	81	68	88	83
	C2	358	367	374	471	63	70	52	61
	C3	14	30	23	31	1	7	5	1
	C4	198	198	194	193	31	31	29	31
	C5	4	4	8	9	1	1	3	1
Q2	C1	595	575	572	479	79	69	86	82
	C2	362	367	378	462	65	69	54	62
	C3	15	30	22	31	1	7	5	1
	C4	199	198	193	194	31	31	29	31
	C5	3	4	9	8	1	1	3	1
Q3	C1	600	574	574	477	78	68	86	81
	C2	360	370	375	467	66	70	54	63
	C3	15	31	26	31	1	7	5	1
	C4	196	195	193	191	31	31	30	31
	C5	3	4	6	8	1	1	2	1
Q4	C1	611	590	538	495	82	76	86	87
	C2	342	356	402	448	54	61	48	50
	C3	13	20	26	23	2	1	4	1
	C4	206	198	204	200	38	39	37	39
	C5	2	10	4	8	1	0	2	0
Q5	C1	624	592	542	498	89	77	88	84
	C2	337	357	403	451	51	62	49	57
	C3	10	22	26	22	2	3	5	1
	C4	200	194	201	195	4	35	34	35
	C5	3	9	2	8	1	0	2 1	80

Table 6.11: Recommender-user stance correlation counts.

user stance bias on both test sets, and TF-IDF primarily on the *random test* set. The knowledge-aware RippleNet, while still leaning towards viewpoints against the given topic, tends to produce recommendations with stance bias scores closer to neutral than the average user stance bias score. However, we observe no statistically significant differences between the user and recommender stance bias in the case of the Transformer and RippleNet recommenders.

We identify five recommender-user stance bias correlation cases:

- *(C1)* Same-direction bias: both the user and the recommender have a tendency toward articles against (in favor) of a particular topic.
- *(C2)* Opposing-direction bias: the user has a preference for articles against (in favor) of a particular topic, while the recommender tends to suggest news with the opposite stance.
- *(C3)* Skewed toward the user: the user has a preference for articles against (in favor) of a particular topic, while the recommender is balanced with regards to the stance of the recommended articles.
- *(C4)* Skewed toward the recommender: the user is balanced with regards to the stance of the read articles, while the recommender tends to suggest articles against (in favor) of a particular topic.
- *(C5)* No bias on either side: neither the user, nor the recommender show any stance bias.

Table 6.11 reports the number of users for each of the five bias cases and questions. We find that cases *(C1)* and *(C2)* dominate, with the former being more likely: recommenders most often reinforce users' stance preferences. The third most common case is *(C4)*, where users are balanced with regards to the stance of articles in their reading history (i.e. average user stance bias score close

	Complete Test Set				Random Test Set			
ID	TF-IDF	Word2vec	Transformer	RippleNet	TF-IDF	Word2vec	Transformer	RippleNet
Q1	0.248 (*)	0.196 (*)	0.187 (*)	-0.022 (-)	0.078 (-)	0.029 (-)	0.162 (**)	0.111 (-)
Q2	0.241 (*)	0.196 (*)	0.176 (*)	-0.008 (-)	0.078 (-)	0.019 (-)	0.148 (**)	0.115 (-)
Q3	0.252 (*)	0.199 (*)	0.195 (*)	-0.009 (-)	0.072 (-)	0.003 (-)	0.143 (-)	0.091 (-)
Q4	0.280 (*)	0.249 (*)	0.143 (*)	-0.019 (-)	0.135 (-)	0.002 (-)	0.230 (*)	0.148 (-)
Q5	0.320 (*)	0.269 (*)	0.168 (*)	-0.011 (-)	0.192 (**)	0.042 (-)	0.230 (*)	0.112 (-)

Table 6.12: Recommender-user stance bias Pearson correlation. Statistical significance is marked with * ($p - value < 0.01$) and (**)($p - value < 0.05$). A '-' denotes no significant difference.

to 0), whereas the recommenders are prone to suggesting articles with a particular viewpoint. Cases *(C3)* and *(C5)* are rare. Overall, the prevalence of *(C1)* suggests that not only do recommenders learn user preferences from their reading history, but also tend to amplify existing viewpoints by recommending articles with the same stances as the ones previously consumed by the user.

Pearson correlations (cf. Table 6.12) further support this picture. All text-based models show significant positive correlation with the existing user stance bias on the *complete test* set for all questions. On the *random test* set, only the Transformer recommender maintains a significant positive correlation with the user bias for four out of five questions, while word2vec shows no statistically significant bias correlation with the user history. The recommendations generated by TF-IDF are significantly positively correlated with the users' stance bias only for question (Q5) ($p - value < 0.05$). RippleNet again shows no statistically significant correlation on either test set, further confirming that the knowledge-aware model is less prone to user bias propagation.

6.5 NeMig: Bilingual News Dataset

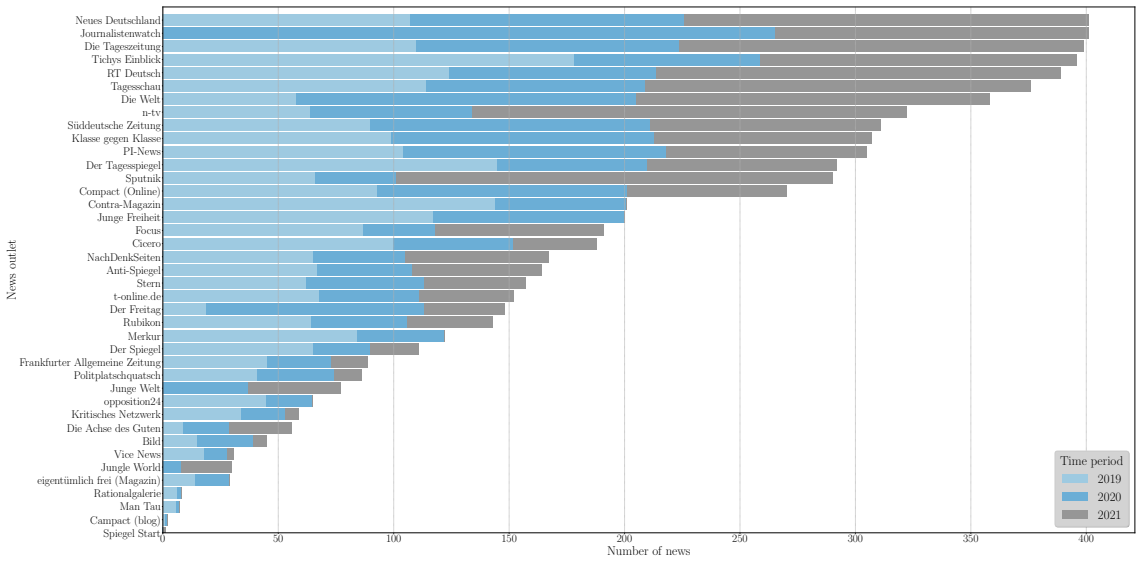
Building on the findings of Section 6.4, we now turn to a broader setting. While GeNeG allowed us to examine recommendation bias in a monolingual scenario, real-world news ecosystems are increasingly multilingual, cross-platform, and interconnected. To investigate whether similar patterns of algorithmic news curation bias occur beyond German-language outlets, and to assess the behavior of more complex news recommenders, we construct an extended multilingual corpora, NeMig, and conduct a new bias analysis on this richer dataset in Section 6.6.

6.5.1 News Data Collection and Annotation

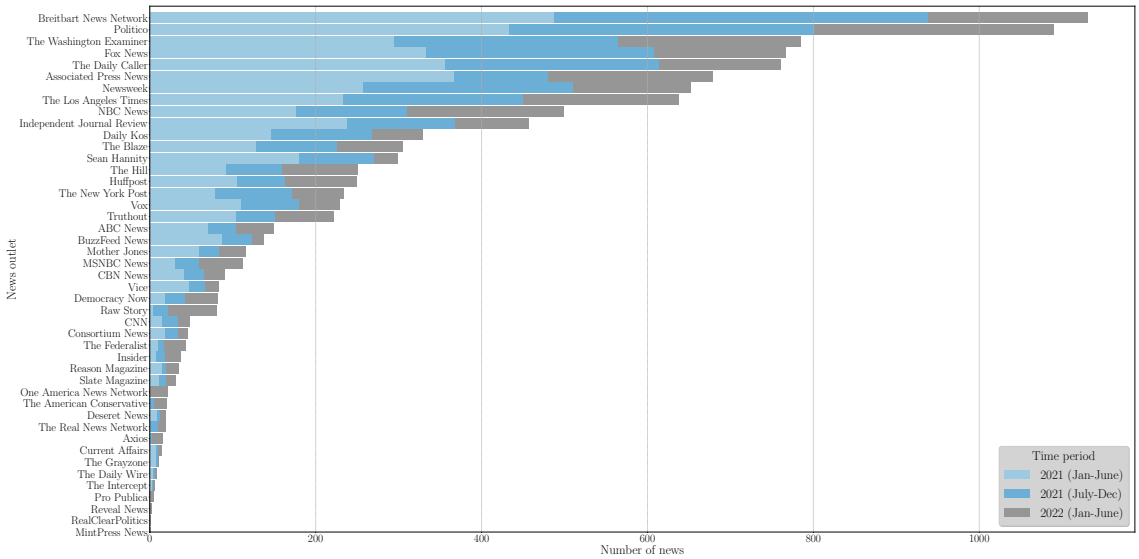
Data Collection. We construct NeMig, a multilingual extension of GeNeG, by crawling news articles from 40 German[6], and, respectively, 45 US[7], legacy and alternative media outlets, selected by a team of media and communication scholars. As in GeNeG, sources span the political spectrum to capture diverse viewpoints on refugees and migration. We update the German news component of GeNeG (initially collected in December 2020), with articles crawled in January 2022. The resulting German corpus includes articles published between January 1st, 2019 and December 31st, 2022. The English corpus comprises US news published between January 1st, 2021 and July 1st, 2022, crawled in August 2022.

To identify migration-related news, we query each outlet for 13 topic-representative keyword stems per language, such as *flüchtl**, *asyl**, or *migration** in German, or *refugee**, *asylum seeker**, or *migrant** in English. We retain only articles written in the target language, with at least 150 words and two keyword stems, filtering out foreign-language content, readers' comments, disclaimers,

[6]For the German news collection, we use the same implementation of the custom web crawlers as in GeNeG.
[7]The implementation of the English news crawlers is available at https://github.com/andreeaiana/us-news



(a) German sebset of NeMig



(b) English subset of NeMig

Figure 6.1: Distribution of news over time and over outlets in the NeMig corpora.

advertisements, and other non-news material. We further discard duplicates, videos, live tickers (i.e., (articles continuously updated with news headlines)), and outliers (i.e., articles deviating by more than two standard deviations from the mean length). Each news item includes the article title, body, and, when available, an abstract, along with metadata such as URL, publishing and modification dates, author names (persons or organizations), and outlet-provided keywords. The final corpora contain 7,346 German and 10,814 English articles. Fig. 6.1 shows the distribution across outlets and over time for the German and English subsets of NeMig. We observe an uneven distribution across publishers, where major outlet dominate, whereas temporal coverage is relatively uniform.

We further develop an automatic annotation and KG construction pipeline to transform NeMig into a richly contextualized resource for studying algorithmic bias and enabling KG-driven recommenders. The annotations capture complementary dimensions of news content, namely ideological context (political leaning), topical structure (sub-topics), affective polarity (sentiment), and real-world references (named entities). The pipeline operates end-to-end, with the exception of the sub-topic

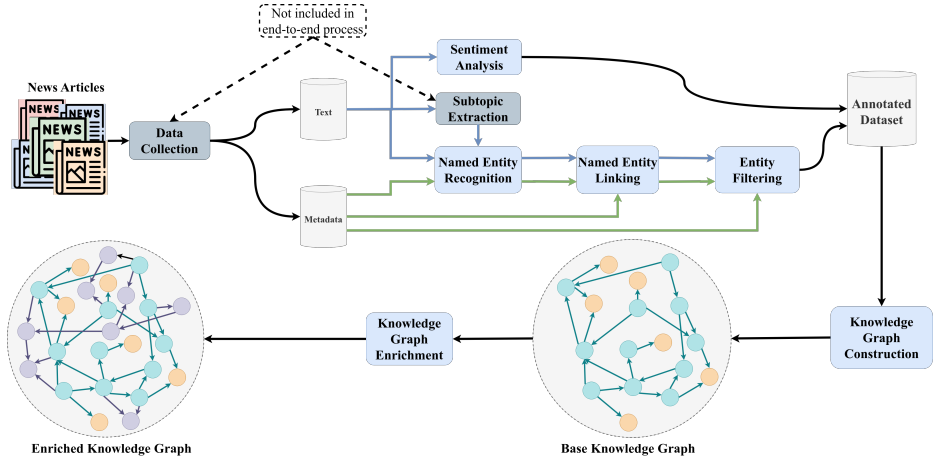


Figure 6.2: Corpora annotation and knowledge graph construction pipeline. Arrow colors denote the applicability of steps to: all components (black), text features (blue), and metadata (green). Node colors indicate: custom or linked resources (blue), literals (orange), and Wikidata neighboring nodes (purple). Edge colors denote properties in the original KG (blue), or extracted from Wikidata (purple).

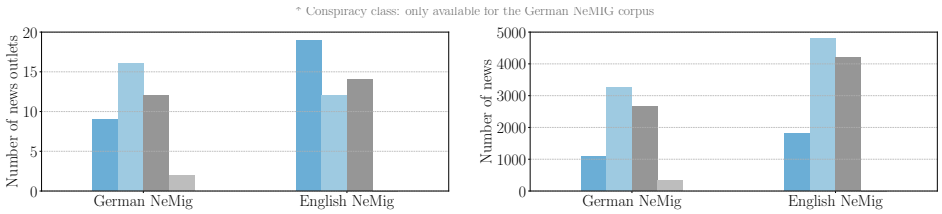


Figure 6.3: Distribution of outlets and news over political orientation classes.

extraction module, which requires manual threshold selection and is therefore, executed separately. A high-level overview of the pipeline is provided in Fig. 6.2.[8]

Political Leaning Annotations. To enable comparative bias analyses in terms of political ideology, we annotate the political orientation of media outlets. We classify US outlets into *left*, *center*, and *right*, according to the AllSides Media Bias Chart[9]. Since no standardized resource exists for German media, the expert team assign outlets to *left*, *center*, *right*, and *conspiracy* categories.[10] Fig. 6.3 shows the distribution over political classes for the German and English subsets of NeMig. While German outlets are predominantly center-aligned, US outlets skew left. At article level, however, the distributions converge more closely across the two language-specific subsets.

Sentiment Annotations. We classify article sentiment using a multilingual XLM-RoBERTa (Conneau et al., 2020), trained on 198M tweets and fine-tuned for sentiment analysis (Barbieri et al., 2022b). We assign each article to *positive*, *neutral*, or *negative* categories. We perfom sentiment classification on the concatenation of article title and abstract, or, if the abstract is missing, the title and first body sentences. Table 6.13 reports sentiment distributions over articles in NeMig. Both corpora contain few positive articles. However, whereas the distribution of negative and neutral articles in the German subset is relatively balanced, the English corpus contains substantially more negative content.

Sub-topic Annotations. Migration-related news discourse spans diverse geopolitical contexts and and

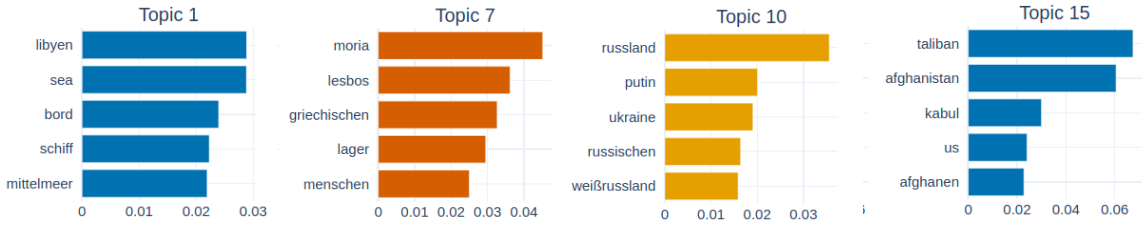
[8]Pipeline implementation and intermediate annotation files are available at https://github.com/andreeaiana/nemig.

[9]https://www.allsides.com/media-bias/media-bias-chart

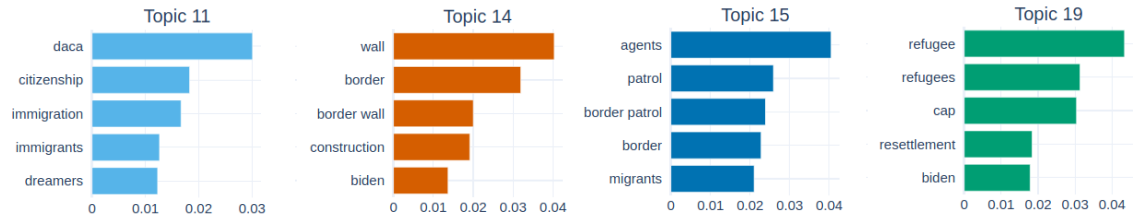
[10]*Man Tau* is excluded from further political orientation bias analysis due to its unclear ideological profile, but its articles remain included in the NeMig corpus.

Dataset	# News		# News / sentiment class			Avg # words		
	Raw	Final	Positive	Negative	Neutral	Title	Abstract	Body
German	8.647	7,346	144	3,476	3,726	7.9	29.4	898.4
English	12,220	10,814	590	2,933	7,291	11.8	24.9	1,013.9

Table 6.13: NeMig corpora statistics.



(a) German news corpus



(b) English news corpus

Figure 6.4: Examples of words assigned to different sub-topics. The X-axis indicates the c-TF-IDF scores of the words representative of each sampled topic.

geographical areas. To extract latent sub-topics from our corpora, we use BERTopic (Grootendorst, 2022), with on a multilingual Sentence Transformer (Reimers and Gurevych, 2020b) for the German news and a pretrained English Sentence Transformer for the English news. We label topics based on the top-most representative terms of the respective cluster of documents, and group topics with fewer than 15 articles into an "other" topic. We extract 25 German and 40 English sub-topics. As seen in Fig. 6.4, topics include migration flows (e.g., Libya, Ukraine), destination contexts (e.g., Greece, US), political crises (e.g., Ukraine's invasion by Russia, US' army withdrawal rom Afghanistan), and policy debates (e.g., US immigration and border control laws).

Named Entity Annotations. We perform named entity recognition (NER) on both article text and metadata using two XLM-RoBERTa (Conneau et al., 2020) LMs, fine-tuned on the German[11] and English[12] subsets of the CoNLL03 dataset (Sang and De Meulder, 2003), respectively, to match the corresponding NeMig corpora. Extracted entities are categorized into persons (PER), organizations (ORG), locations (LOC), and miscellaneous (MISC). For textual content (e.g., title, abstract, article body), we apply sentence segmentation before NER and retain only the entities recognized by the model. For metadata, we input author names, outlet-provided keywords, and sub-topic labels; if no entity is detected for a given term, the original string is retained. In the KG construction phase (see § 6.5.2), we incorporate both successfully recognized entities and additional metadata signals (e.g., frequent keywords, unrecognized authors) to model relationships between articles.

Named Entity Disambiguation. We next disambiguate the extracted entities via entity linking (EL) to Wikidata (Vrandečić and Krötzsch, 2014) using a multilingual sequence-to-sequence linker (De Cao

[11] https://huggingface.co/xlm-roberta-large-finetuned-conll03-german
[12] https://huggingface.co/xlm-roberta-large-finetuned-conll03-english

Component		German			English		
		Total	Linked	Not linked	Total	Linked	Not linked
Text	Title	1,075	1,075	0	1,542	1,542	0
	Abstract	2,383	2,383	0	2,365	2,365	0
	Body	19,683	19,683	0	33,857	33,857	0
Metadata	Publishers	40	34	6	45	44	1
	Authors	490	193	297	1,242	386	856
	Keywords	4,481	2,395	2,086	5,175	2,636	2,539

Table 6.14: Statistics of extracted named entities.

et al., 2022; Liu et al., 2020b). The input to the linked consists of text sequences containing a single named entity annotated with boundary tokens, taken either from the article text (e.g., title, abstract, sentence-segmented news body) or from metadata fields (e.g., metadata-derived named entities, original outlet-provided keyword, author names). The model outputs the Wikidata entity, its QID, language label, and a confidence score. We select Wikidata as the external knowledge base for EL and graph augmentation due to its open access, wide coverage, and recency (Heist et al., 2020) – an important attribute for the fast-evolving news domain.

Entity Filtering. The previous step can produce errors, including entities that exist in Wikidata but were not linked (e.g., the keyword *academia* corresponds to QID *Q1211427*), as well as incorrect link assignments. We thus apply a multi-stage filtering procedure to remove incorrectly extracted or linked entities. For named entities originating from textual components, we: (1) apply a confidence threshold on the NER model predictions; (2) exclude entities without a Wikidata page or those linked to Wikidata disambiguation pages; (3) discard entities whose type does not match the Wikidata entity type, based on type-specific properties[13] (4) apply a confidence threshold on the EL model; and (5) remove entities whose Wikidata label appears only once in the corpus and is in a foreign language. For authors and keywords, we use similar filtering pipelines. Table 6.14 shows the statistics for the named entity recognition and linking steps.

6.5.2 Knowledge Graph Construction

Wikidata (Vrandečić and Krötzsch, 2014) is a large-scale knowledge graph widely used t model and integrate real-world and common-sense knowledge into downstream applications, including recommender systems (Guo et al., 2020b; Iana et al., 2024). However, not all Wikidata information is relevant for the news recommendation domain. We therefore construct a domain-specific KG from NeMig, integrating textual and metadata-derive information with Wikidata triples of the extracted named entities and their k-hop neighbors.

Base Graph Construction. NeMigKG is a heterogeneous KG in which nodes correspond to news content elements and real-world entities (e.g., authors, locations), and edges represent relations among them. Fig. 6.5 provides an example of an article represented in NeMigKG.

Relations. Relations in NeMigKG are modeled using established schema and vocabularies. We represent edges between articles and their content or metadata using *schema.org*[14] predicates, and those involving extracted named entities using the *Simple Event Model* (Van Hage et al., 2011). We additionally capture miscellaneous named entities via *schema:mentions*, and map the provenance of named entities to textual fields using *dcterms:isReferencedBy*.[15]. We further introduce two custom

[13]We use the following properties to infer entity type: PER is *instance of human*; ORG has *headquarters location*, *inception*, or is *founded by*; LOC has *coordinate location*.

[14]https://schema.org/

[15]https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/isReferencedBy/

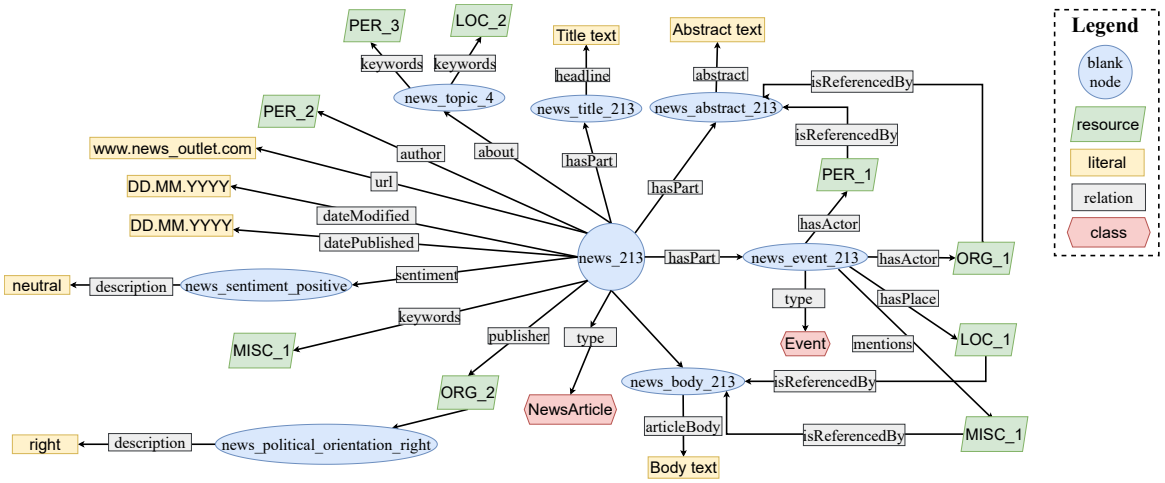


Figure 6.5: Example of an article's representation in NeMigKG. The node colors indicate: (i) blank nodes (blue); (ii) disambiguated or custom resources (green); (iii) literals (yellow); (iv) classes (red). The gray boxes denote relation types.

relations, `nemig:sentiment` and `nemig:political_orientation`, to link news to sentiment labels and media outlets to political-leaning labels, respectively. Finally, we detect pairs of related articles, identified by body overlap ratio and identical titles with different publishers, and connect them with *schema:isBasedOn*. Typically, the newer article represents an update, extension, or more detailed discussion of the earlier one.

Nodes. NeMigKG includes two categories of nodes: (i) *literals* (e.g., titles, publication dates), and (ii) *resources*, representing named entities extracted from article content (e.g. persons, locations) or metadata (i.e., publisher, author). Resource nodes are further divided into: *Wikidata resources* – entities that were successfully disambiguated and linked to Wikidata; and *custom resources* – entities not present in or not linkable to Wikidata but still informative (e.g., common keywords or authors absent from Wikidata). We represent each article in NeMigKG with a unique identifier, and create additional nodes for each title, abstract, body, topic, sentiment, political leaning, and event mapped in the KG, as exemplified in Fig. 6.5.

Extended Graph Construction. To enrich NeMigKG with broader factual context, we extend the graph with up to *2*-hop neighbors from Wikidata for all existing *Wikidata resources.* For each *Wikidata resource*, we first retrieve all triples with another entity as the tail (ignoring literal values), and iteratively add both the tail entities and their triples until the two-hop neighborhood is exhausted. After enrichment, we prune sink nodes (i.e., nodes with degree smaller than two), to remove isolated or uninformative nodes that cannot contribute to relationa reasoning.

Graph types. Since many KG embedding methods ignore literal nodes, we generate four variants of NeMigKG for each language: (i) *Base* – contains all literals and resource nodes from the news corpus; (ii) *Entities* – contains resource nodes only and no literals; (iii) *Enriched entities* – contains resource nodes plus corresponding Wikidata k-hop triples; (iv) *Complete* – constitutes the enriched entities graph with literals reintroduced. Table 6.15 reports the statistics for each graph variant.

6.5.3 User Data Collection

We collect user data in terms of (i) explicit click feedback, (ii) demographics, and (iii) political information for both language settings (German and English) to complement the annotated news corpus and enable analysis of algorithmic news curation on recommender sytem users. For each corpus, we conduct an online user study to measure political polarization effects in news recommendation. Participants are recruited via certified online access panels, using quota sampling to approximate the

		German NeMigKG					English NeMigKG				
		Base	Entities	Enriched ($k=1$)	Enriched ($k=2$)	Complete	Base	Entities	Enriched ($k=1$)	Enriched ($k=2$)	Complete
Triples	Nodes	89,601	59,480	97,021	367,514	397,635	134,775	91,833	138,734	433,965	476,907
	Relations	19	11	768	1,172	1,180	19	11	828	1,195	1,203
	Triples	516,948	458,482	821,529	2,846,849	2,905,315	933,801	8 47,695	1,354,454	3,599,561	3,685,667
Nodes	% resources	66.38	100	100	100	92.42	68.14	100	100	100	91.00
	% literals	33.62	0	0	0	7.58	31.86	0	0	0	9.00
Resources	% blank	61.28	61.29	37.57	9.92	9.92	58.48	58.48	38.71	12.38	12.38
	% custom (not linked)	4.00	4.00	2.45	0.65	0.65	3.68	3.68	2.43	0.78	0.78
	% Wikidata (linked)	34.71	34.71	59.97	89.43	89.43	37.84	37.84	58.86	86.85	86.85

Table 6.15: Statistics of the German and English NeMigKG.

demographic composition of Internet users in Germany and the United States, respectively, aged 18 to 74. The German cohort comprises 3,432 participants (50.2% male, 49.6% female), of whom 44.6% hold an *Abitur* (e.g., German university entrance certificate). The US cohort includes 3,000 users (46.2% male, 52.5% female), with 33.2% holding at least a high-school diploma.

The study consists of two phases. In the initial phase, each participant rates 10 articles as *worthy* or *not worthy of further reading*. These explicit preferences form the user's *Click History*, which serves as the basis for constructing user profiles. In the impression phase, participant receive five rounds of six recommended articles each (e.g., impressions) and again provide a binary click feedback. These interactions constitute the *Impression Log*. Following Wu et al. (2020e), we structure the anonymized behavioral data as *[impression ID, user ID, Click History, Impression Log]*, where the *Impression Log* contains the IDs of the news shown to the user and the label indicating whether the user clicked on them.

To contextualize behavioral responses, participants complete pre- and post-exposure questionnaires. Prior to the stimulus phase, we collect data on media use (*MED[1-9]*), political attitudes (*POL[1-9]*), and empathy levels (*EMP[1-8]*). After exposure to the news recommenders, we gather socio-demographic characteristics (e.g., gender, age, education, nationality, migration background, household income), self-reported emotional responses (*EMO[1-14]*), and measure the ideological (*IPO[1-12]*), affective (i.e., party identification) and perceived (*PPO*[1-6]) polarization, political participation (*PPA[1-8]*), and prosocial attitudes (*PRO[1-5]*).

Interaction density differs across studies: the German dataset has a sparsity of only 0.02% (i.e., fewer than 1% of presented articles remain unseen by any participant), whereas the US dataset is sparser, with 9.27% of articles never interacted with by users. Table 6.16 presents an excerpt of the questionnaire used in the US study. The complete questionnaires, including answer scales and documentation, are publicly available on the project repository and Zenodo (Iana et al., 2022a).

6.5.4 Data Availability

NeMig is released in two formats: (i) we provide the complete KG (including news bodies) in N-Triples format under restricted access on Zenodo (Iana et al., 2022a); the full text of articles is provided upon request to comply with copyright restrictions, and (ii) an open version containing all features except article bodies is available under a CC BY-NC-SA 4.0 license in tabular format on GitHub.[16] The anonymized user-interaction and survey datasets are likewise available on both Zenodo (Iana et al., 2022a) and GitHub.[17]

[16] https://github.com/andreeaiana/nemig_nrs/data
[17] https://github.com/andreeaiana/nemig/user_data

Category	Variable	Question (Answer / Answer Scale)	
Demographics	Gender	Please indicate your gender.	
		0 = male, 1 = female , 2 = other, 3 = no answer	
	Nationality	What is your citizenship?	
		0 = U.S. citizenship, 1 = U.S. and another non-U.S. citizenship, 2 = Only non-U.S. citizenship, 3 = No Answer	
Empathy		How strongly do you agree with the following statements?	
	EMP1	When someone else is feeling excited, I tend to get excited too.	7-point Likert scale
	EMP2	It upsets me to see someone being treated disrespectfully	(1=Strongly disagree to 7=Strongly agree)
Ideological Polarization		In the following, you will see a series of opposing statements. Please indicate how strongly you agree or disagree with the statements. There are no right or wrong answers.	
	IPO1	The U.S. should take in more refugees.	
	IPO2	The U.S. has already taken in too many refugees.	7-point Likert scale
	IPO5	Immigrants strive for peaceful cohabitation with U.S.-Americans.	(1=Strongly disagree to 7=Strongly agree)
	IPO6	Immigrants are hostile toward U.S.-Americans.	
Emotions		Which emotions do you feel towards refugees and immigrants in the USA?	
	EMO1	Anger	
	EMO2	Fear	7-point Likert scale
	EMO9	Compassion	(1=Strongly disagree to 7=Strongly agree)
	EMO13	Gratitude	
Media Usage		I receive information about US-American politics via:	
	MED1	newspapers and magazines or their websites (e.g. New York Times, The Wallstreet Journal, ...)	
	MED2	TV networks or their websites (e.g. Fox News, CNN...)	7-point Likert scale
	MED4	Facebook (for political information)	(1=Never to 7=Very Often)
Perceived Polarization		How strongly do you agree or disagree with the following statements?	
	PRO1	Democratic and Republican partisans in the U.S. are increasingly hostile to one another.	7-point Likert scale
	PRO3	Democratic and Republican partisans in the U.S. are very polarized.	(1=Strongly disagree to 7=Strongly agree)
Affective Polarization		In the following we would like to know about your party identification. Please mark on the scale how warm or cold you feel towards the respective parties.	
	REP	Republican	0 (negative) to 100 (positive)
	DEM	Democrat	
Political Scale		Where on the scale would you place your political point of view?	
	POL1	11-point Likert Scale (1 = Left, 6 = Center, 11 = Right)	
Political Topics		Please indicate how strongly you agree or disagree with the following statements.	
	POL2	I am generally very interested in politics.	7-point Likert scale
	POL3	I regularly inform myself about current political affairs in the U.S.	(1=Strongly disagree to 7=Strongly agree)
Participation		Please indicate how likely it is that you will engage in the following activities in the near future	
	PPA1	Participating in an online political discussion on the topic of immigration to the U.S.	7-point Likert scale
	PPA6	Participating in a political demonstration on the topic of immigration to the U.S.	(1=Not likely at all to 7=Very likely)
Prosocial Behavior		Please indicate how strongly you agree or disagree with the following statements.	
	PRO3	I am willing to make a one-time donation of __$ for refugees in the U.S.	float number >= 0
	PRO4	How often do you have professional contact (e.g. at work or at school) with immigrants?	7-point Likert scale (1 = Never to 7 = Very often)

Table 6.16: Excerpt of questionnaire used in the US online user study.

6.6 Bias Analyses Revisited

We analyze the performance of a range of NNRs on NeMig across multiple evaluation axes. This setup extends the earlier experiments conducted on GeNeG (see Section 6.4) by moving from simpler baselines to state-of-the neural architectures, enabling us to study how increased model expressiveness influences both personalization performance and bias-related behavior.

6.6.1 Experimental Setup

Recommendation Models. We benchmark several NNRs with different combinations of news and user encoders.[18] As introduced in Section 3.1, news encoders convert input features (e.g., title, topics, entities) into a news embedding by contextualizing pretrained word embeddings (Pennington et al., 2014), whereas user encoders aggregate clicked-news embeddings into a user-level representation.

- *NRMS* (Wu et al., 2019e) embeds news titles with a two-layer encoder consisting of multi-head self-attention (Vaswani et al., 2017), followed by additive attention (Bahdanau et al., 2015); a similar architecture is used to learn user representations.

- *NAML* (Wu et al., 2019a) uses a sequence of CNN (Kim, 2014) and additive attention to learn representations from news titles and abstracts, additionally leveraging categories through a linear category encoder; its user encoder consists of an additive attention layer.

[18]The implementation of the news recommenders is available at https://github.com/andreeaiana/nemig_nrs.

- *MINS* (Wang et al., 2022c) embeds the titles and abstracts of news with the same encoder an NRMS encoder (Wu et al., 2019e), and categories through a linear embedding layer; it learns user representations with a combination of multi-head self-attention, multi-channel GRU-based recurrent network (Cho et al., 2014), and additive attention.
- *CAUM* (Qi et al., 2022b) combines an NRMS-style text encoder (Wu et al., 2019e) with an entity embedder composed of attention layers in order to learn news representations; it produces candidate-aware user representations with a candidate-aware self-attention network which models long-range dependencies between clicked news, conditioned on the candidate, combined with a candidate-aware CNN that captures short-term user interests from adjacent clicks, also conditioned on the candidate's content.
- *DKN* (Wang et al., 2018b) is a knowledge-aware recommender which learns representations from news titles with a knowledge-aware CNN (Kim, 2014) over aligned embeddings of words and entities, and of users with a candidate-aware attention network.
- *TANR* (Wu et al., 2019c) follows NAML's architecture (Wu et al., 2019a), but omits explicit category embeddings while jointly optimizing the model for topic classification alongside content personalization.
- *SentiDebias* (Wu et al., 2022h), builds on NRMS (Wu et al., 2019e) and uses adversarial learning to reduce sentiment bias in user representations and to generate sentiment-diverse recommendations.

Evaluation Metrics. We evaluate the recommenders along two complementary axes: (i) traditional accuracy-based performance and (ii) aspect-based diversity and personalization. For the content personalization, we report standard metrics: AUC, MRR, NDCG@k ($k \in [3, 6]$.

We extend the preliminary analyses on GeNeG (cf. Section 6.4), which relied on average-aggregate measure, and now adopt distribution-based metrics. This shift allows us to capture how the recommendation list reflects the distribution of a given categorical article attribute, referred to as *aspect* A_p, such as topic, sentiment, stance, or political leaning. These metrics quantify whether models prioritize *varied exposure (diversification)* or *preference alignment (personalization)* with the user's historical reading patterns. We experiment with three aspects: *topical categories (ctg)*, the *sentiment (snt)*, and the *political leaning (pol)* of news. Note that we use the political leaning of a news outlet as proxy for inferring the political leaning of an article originating from that outlet.

We define *aspect-based diversification* as the level of uniformity of an aspect's distribution among the recommended news. We compute aspect-based diversity of recommendations at position k as the normalized entropy of the distribution of aspect A_p's values in the recommendation list:

$$D_{A_p}@k = -\sum_{j \in A_p} \frac{p(j) \log p(j)}{\log(|A_p|)} \tag{6.2}$$

where $A_p \in \{ctg, snt, pol\}$, and $|A_p|$ is the number of A_p classes.

Conversely, we define *aspect-based personalization* as the level of homogeneity between a user's recommendations and clicked news w.r.t. the distribution of an aspect (e.g., sentiment). Higher values indicate stronger alignment. We evaluate aspect-based personalization using the generalized Jaccard similarity (Bonnici, 2020):

$$PS_{A_p}@k = \frac{\sum_{j=1}^{|A_p|} \min(\mathcal{R}_j, \mathcal{H}_j)}{\sum_{j=1}^{|A_p|} \max(\mathcal{R}_j, \mathcal{H}_j)}, \tag{6.3}$$

where R_j and H_j represent the probability of a news with class j of A_p to be contained in the recommendations list R, and, respectively, in the user history H. All metrics are bounded in $[0, 1]$.

	German				English			
Model	AUC	MRR	NDCG@3	NDCG@6	AUC	MRR	NDCG@3	NDCG@6
NRMS	$57.96_{\pm 0.66}$	$47.83_{\pm 0.46}$	$44.90_{\pm 0.65}$	$60.49_{\pm 0.35}$	$52.40_{\pm 0.89}$	$41.46_{\pm 0.59}$	$37.49_{\pm 1.07}$	$54.21_{\pm 0.53}$
NAML	$50.49_{\pm 0.13}$	$47.37_{\pm 0.84}$	$44.27_{\pm 1.20}$	$60.13_{\pm 0.65}$	$50.02_{\pm 0.20}$	$41.82_{\pm 0.70}$	$38.14_{\pm 0.84}$	$54.49_{\pm 0.62}$
MINS	$57.70_{\pm 0.53}$	$47.85_{\pm 0.78}$	$44.96_{\pm 0.33}$	$60.50_{\pm 0.71}$	$52.91_{\pm 0.68}$	$41.96_{\pm 1.13}$	$38.19_{\pm 0.41}$	$54.59_{\pm 0.41}$
CAUM	$58.18_{\pm 0.85}$	$48.17_{\pm 1.08}$	$45.27_{\pm 0.60}$	$60.74_{\pm 0.55}$	$53.10_{\pm 0.78}$	$41.83_{\pm 0.93}$	$38.16_{\pm 1.42}$	$54.50_{\pm 0.32}$
DKN	**$58.73_{\pm 0.39}$**	**$48.20_{\pm 0.39}$**	**$45.48_{\pm 0.43}$**	**$60.78_{\pm 0.29}$**	**$53.69_{\pm 0.66}$**	**$42.14_{\pm 0.66}$**	**$38.57_{\pm 0.83}$**	**$54.74_{\pm 0.55}$**
TANR	$50.62_{\pm 0.14}$	$47.28_{\pm 0.44}$	$44.22_{\pm 0.51}$	$60.06_{\pm 0.33}$	$50.17_{\pm 0.10}$	$41.82_{\pm 0.12}$	$38.26_{\pm 0.33}$	$54.49_{\pm 0.17}$
SentiDebias	$56.80_{\pm 0.30}$	$47.27_{\pm 0.44}$	$44.09_{\pm 0.18}$	$60.04_{\pm 0.44}$	$52.63_{\pm 0.90}$	$41.63_{\pm 1.34}$	$37.58_{\pm 0.13}$	$54.33_{\pm 0.58}$

Table 6.17: Recommendation performance of different NNRs. The best results per column are highlighted in bold, the second best are underlined.

Training and Optimization Details. We use pretrained 300-dimensional German[19] and English[20] GloVe embeddings (Pennington et al., 2014) and 100-dimensional TransD embeddings (Ji et al., 2015) pretrained on NeMigKG to initialize the word and entity embeddings of the NNRs. Note that we use the entity embeddings trained on the variant of NeMigKG enriched with 1-hop Wikidata neighbors. Moreover, we use the extracted sub-topics as topical categories in NAML, MINS, and TANR, and embed them with 100-dimensional vectors. Following Wu et al. (2022b), we sample four negatives per positive example. We train all models for 10 epochs, using a batch size of 4.[21] We optimize with the Adam algorithm (Kingma and Ba, 2014), with the learning rate set to 1e-5. We set all other model-specific hyperparameters to optimal values reported in the respective papers. We use 70% randomly sampled user interactions for training the recommenders, 10% as validation data, and the remaining 20% as test set. We repeat each experiment five times with different random seeds and report averages and standard deviations. We normalize the scores to the [0, 100] range.

6.6.2 Results and Discussion

We first discuss the recommendation performance of the aforementioned NNR. Next, we analyze whether they are prone to implicit biases across three dimensions: *political leaning*, *topical categories*, and *sentiment orientation*. Finally, we investigate the contribution of different input features to the quality of NeMigKG, as reflected in the downstream performance of knowledge-aware recommenders.

Content Personalization. Table 6.17 reports the content personalization performance for both datasets. Despite the relatively limited dataset size, the state-of-the-art neural models demonstrate strong predictive performance. We observe highly similar results between NAML and TANR, as well as between NRMS and SentiDebias. Each pair of models shares nearly identical user and news encoder architectures, differing primarily in their secondary training objectives. This suggests that auxiliary objectives, such as topic prediction in TANR or sentiment debiasing in SentiDebias, exert minimal influence on pure recommendation accuracy.

Overall, the two knowledge-aware models (DKN and CAUM) outperform the other recommenders, indicating that incorporating structured external knowledge enables the detection of additional content-level relationships and leads to richer news representations. These findings partially diverge from those in Chapter 4, where DKN underperformed while CAUM outperformed the other NNRs. However, that analysis relied on the much larger MIND dataset (Wu et al., 2020e). Our results therefore suggest that both knowledge-awareness and candidate-awareness user modeling are particularly beneficial in data-scarce scenarios. We also find stronger content personalization performance for the German dataset, a difference which we attribute to the higher sparsity of the English dataset, where a substantially larger fraction of articles are unseen by users.

[19] https://www.deepset.ai/german-word-embeddings
[20] https://nlp.stanford.edu/projects/glove/
[21] We trained each model on a single NVIDIA RTX 2080Ti GPU.

	German				English			
Model	NDCG@3	D_{ctg}@3	D_{snt}@3	D_{pol}@3	NDCG@3	D_{ctg}@3	D_{snt}@3	D_{pol}@3
NRMS	$44.90_{\pm 0.65}$	$18.55_{\pm 0.25}$	$33.75_{\pm 0.54}$	$30.68_{\pm 0.32}$	$37.49_{\pm 1.07}$	$22.19_{\pm 0.34}$	$32.50_{\pm 0.27}$	$26.31_{\pm 0.94}$
NAML	$44.27_{\pm 1.20}$	**$19.31_{\pm 0.21}$**	$33.63_{\pm 0.18}$	$30.41_{\pm 0.37}$	$38.14_{\pm 0.84}$	**$23.89_{\pm 0.65}$**	$32.08_{\pm 0.64}$	$25.66_{\pm 1.24}$
MINS	$44.96_{\pm 0.33}$	$18.18_{\pm 0.44}$	$33.79_{\pm 0.34}$	$30.10_{\pm 0.44}$	$38.19_{\pm 0.41}$	$21.94_{\pm 1.11}$	$32.34_{\pm 0.16}$	$25.22_{\pm 1.11}$
CAUM	$45.27_{\pm 0.60}$	$18.09_{\pm 0.33}$	$33.13_{\pm 0.43}$	$30.42_{\pm 0.33}$	$38.16_{\pm 1.42}$	$21.90_{\pm 0.63}$	$32.09_{\pm 0.26}$	$25.90_{\pm 0.63}$
DKN	**$45.48_{\pm 0.43}$**	$18.46_{\pm 0.16}$	$33.79_{\pm 0.44}$	$30.92_{\pm 0.59}$	**$38.57_{\pm 0.83}$**	$22.04_{\pm 0.40}$	$32.46_{\pm 0.66}$	**$26.58_{\pm 0.90}$**
TANR	$44.22_{\pm 0.51}$	$18.52_{\pm 0.18}$	$33.70_{\pm 0.36}$	$30.39_{\pm 0.38}$	$38.26_{\pm 0.33}$	$21.59_{\pm 0.19}$	$32.41_{\pm 0.68}$	$26.47_{\pm 0.82}$
SentiDebias	$44.09_{\pm 0.18}$	$18.66_{\pm 0.30}$	**$33.84_{\pm 0.34}$**	**$30.99_{\pm 0.30}$**	$37.58_{\pm 0.13}$	$22.10_{\pm 1.11}$	**$33.26_{\pm 0.34}$**	$25.81_{\pm 1.11}$

Table 6.18: Content personalization and aspect diversity (in terms of topical categories, sentiments, and political leaning) performance of different NNRs. The best results per column are highlighted in bold, the second best are underlined.

	German				English			
Model	NDCG@3	PS_{ctg}@3	PS_{snt}@3	PS_{pol}@3	NDCG@3	PS_{ctg}@3	PS_{snt}@3	PS_{pol}@3
NRMS	$44.90_{\pm 0.65}$	$20.72_{\pm 0.45}$	**$42.66_{\pm 0.30}$**	$36.55_{\pm 0.09}$	$37.49_{\pm 1.07}$	$18.48_{\pm 0.18}$	$41.38_{\pm 0.40}$	$40.66_{\pm 0.22}$
NAML	$44.27_{\pm 1.20}$	$20.42_{\pm 0.31}$	$42.59_{\pm 0.25}$	$36.23_{\pm 0.32}$	$38.14_{\pm 0.84}$	$18.24_{\pm 0.81}$	**$41.43_{\pm 0.19}$**	**$41.17_{\pm 0.30}$**
MINS	$44.96_{\pm 0.33}$	**$21.10_{\pm 0.60}$**	$42.56_{\pm 0.34}$	**$36.74_{\pm 0.27}$**	$38.19_{\pm 0.41}$	$18.83_{\pm 0.41}$	$41.20_{\pm 0.16}$	$41.15_{\pm 0.21}$
CAUM	$45.27_{\pm 0.60}$	$20.85_{\pm 0.79}$	$42.48_{\pm 0.43}$	$36.59_{\pm 0.36}$	$38.16_{\pm 1.42}$	$18.67_{\pm 1.45}$	$41.41_{\pm 0.26}$	$40.95_{\pm 0.41}$
DKN	**$45.48_{\pm 0.43}$**	$20.95_{\pm 0.35}$	$42.51_{\pm 0.36}$	$36.08_{\pm 0.24}$	**$38.57_{\pm 0.83}$**	$18.56_{\pm 0.35}$	$41.25_{\pm 0.14}$	$40.70_{\pm 0.15}$
TANR	$44.22_{\pm 0.51}$	$20.88_{\pm 0.24}$	$42.64_{\pm 0.19}$	$36.42_{\pm 0.34}$	$38.26_{\pm 0.33}$	**$19.10_{\pm 0.29}$**	$41.27_{\pm 0.24}$	$40.98_{\pm 0.20}$
SentiDebias	$44.09_{\pm 0.18}$	$20.50_{\pm 0.39}$	$42.64_{\pm 0.34}$	$36.40_{\pm 0.24}$	$37.58_{\pm 0.13}$	$18.57_{\pm 0.33}$	$41.29_{\pm 0.34}$	$40.84_{\pm 0.20}$

Table 6.19: Content and aspect personalization (in terms of topical categories, sentiments, and political leaning) performance of different NNRs. The best results per column are highlighted in bold, the second best are underlined.

Aspect-based Diversity and Personalization. We next examine aspect-based diversity and personalization across topical categories (*ctg*), sentiment (*snt*), and political leaning (*pol*). We report content personalization (NDCG@k), aspect-based diversity (D_{ctg} for topical categories, D_{snt} for sentiment polarization, and D_{pol} for political leaning), and aspect-based personalization (PS_{ctg}, PS_{snt}, and PS_{pol}) for $k = 3$.

Table 6.18 summarizes the results on content personalization and aspect diversity. As expected, SentiDebias (Wu et al., 2022h) achieves the highest sentiment diversity across both datasets. NAML (Wu et al., 2019a) yields the most topically diverse recommendations by a large margin. The latter result is somewhat surprising and contrasts with observations on other datasets (see Chapter 7), given that NAML leverages topical information to strengthen alignment to users' preferences rather than promote diversity. In contrast, political-aspect diversity is nearly uniform across models, likely because none explicitly target diversification with regards to this aspect. Consistent with Chapter 7, we observe a negative correlation between aspect-based diversity and personalization (cf. Table 6.19): improving diversity typically reduces personalization. We observe that categorical and political personalization align more closely with content-level personalization than sentiment personalization – an intuitive result, as users' reading choices are more strongly influences by topical interests and political orientation than sentiment polarity,

Knowledge Graph Ablations. The type of data contained in the knowledge graph is paramount for training high quality knowledge graph embeddings, which are further exploited by knowledge-aware recommenders to improve the accuracy of predictions. To evaluate the contribution of different input features, we pretrain entity embeddings using the TransD (Ji et al., 2015) method on several variants of NeMigKG and assess their downstream impact on the performance of DKN. Table 6.20 reports the results in terms of NDCG@k and aspect-based diversity D_{A_p}. Augmenting NeMigKG with topical,

	German					English				
NeMigKG	AUC	NDCG@3	D_{ctg}@3	D_{snt}@3	D_{pol}@3	AUC	NDCG@3	D_{ctg}@3	D_{snt}@3	D_{pol}@3
content entities	$58.72_{\pm 0.45}$	$45.57_{\pm 0.44}$	$18.44_{\pm 0.16}$	$33.67_{\pm 0.38}$	$31.17_{\pm 0.56}$	$53.67_{\pm 0.63}$	$38.60_{\pm 1.01}$	$21.99_{\pm 0.45}$	$32.33_{\pm 0.34}$	$26.71_{\pm 0.75}$
+ topical categories	$58.82_{\pm 0.23}$	$45.74_{\pm 0.31}$	$18.43_{\pm 0.09}$	$33.83_{\pm 0.39}$	$30.93_{\pm 0.64}$	$53.67_{\pm 0.50}$	$38.68_{\pm 0.72}$	$22.02_{\pm 0.38}$	$32.20_{\pm 0.45}$	$26.47_{\pm 1.12}$
+ sentiment	$58.81_{\pm 0.26}$	$45.69_{\pm 0.38}$	$18.48_{\pm 0.08}$	$33.77_{\pm 0.44}$	$30.98_{\pm 0.65}$	$53.70_{\pm 0.53}$	$38.53_{\pm 0.73}$	$21.95_{\pm 0.37}$	$32.17_{\pm 0.18}$	$26.89_{\pm 0.72}$
+ political leaning	$58.73_{\pm 0.42}$	$45.55_{\pm 0.43}$	$18.45_{\pm 0.16}$	$33.68_{\pm 0.47}$	$31.00_{\pm 0.45}$	$53.73_{\pm 0.72}$	$38.66_{\pm 0.79}$	$22.00_{\pm 0.46}$	$32.28_{\pm 0.36}$	$26.74_{\pm 0.52}$
+ 1-hop neighbors	$58.82_{\pm 0.33}$	$45.74_{\pm 0.46}$	$18.50_{\pm 0.09}$	$33.77_{\pm 0.40}$	$31.20_{\pm 0.50}$	$53.78_{\pm 0.44}$	$38.76_{\pm 0.65}$	$21.95_{\pm 0.37}$	$32.16_{\pm 0.25}$	$26.60_{\pm 0.59}$
+ 2-hop neighbors	$58.73_{\pm 0.39}$	$45.48_{\pm 0.43}$	$18.46_{\pm 0.16}$	$33.79_{\pm 0.44}$	$30.92_{\pm 0.59}$	$53.69_{\pm 0.66}$	$38.57_{\pm 0.83}$	$22.04_{\pm 0.40}$	$32.46_{\pm 0.66}$	$26.58_{\pm 0.90}$

Table 6.20: Content personalization and aspect-based diversity (in terms of topical categories, sentiment, and political leaning) performance of DKN with different variants of NeMigKG.

sentiment, and political information increases aspect-level diversity while having minimal effects on accuracy-based performance. Further enriching the graph with external Wikidata neighbors is beneficial up to one hop, but incorporating two-hop information reduces performance. This suggests that moderate contextualization via external knowledge is helpful in data-scarce settings, whereas excessive augmentation may dilute domain-specific signals and negatively affect recommendation quality.

6.7 Conclusion

In this Chapter, we introduced GeNeG, a German news corpus on the topic of refugees and migration, enriched with sentiment polarity and stance annotations Using this resource, we conducted a preliminary analysis of news recommender bias, showing that text-based models tend to amplify user attitudes with respect to the articles' sentiment and stance. Such amplification risks reinforcing filter-bubble dynamics by prioritizing emotionally or ideologically aligned content. These results echo findings from prior work by Gao et al. (2018b) on the effects of stance and credibility labels on online news selection and consumption; such labels intensify users' inclination to consume more belief-congruent content, potentially increasing their susceptibility to polarized views and misinformation. In contrast, a knowledge-aware recommender relying on entity representations exhibited reduced sentiment and stance bias, though at the expense of pure click-prediction performance.

Building on these insights, we extended GeNeG into NeMig, a bilingual (German-English) news collection paired with user interaction data and knowledge graphs linked to Wikidata. NeMig expands the scope of analysis by incorporating political leaning and topic annotations, enabling multidimensional evaluation across sentiment, political orientation, and topical diversity. Our experiments on NeMig confirm that knowledge-enhanced models can provide benefits both in accuracy, particularly in data-scarce settings, and in reducing certain forms of aspect-based bias.

Taken together, these datasets and findings underscore the importance of evaluating recommender systems not only on accuracy, but also on their impact on exposure diversity and potential polarization. Our results highlight a key tension between accuracy-driven personalization and aspect-based diversity in algorithmic news curation. Looking ahead, several research directions emerge. First, while this Chapter focuses on content-based algorithms, future studies should examine whether collaborative-filtering approaches exhibit similar bias patterns. Second, our analysis primarily captures political bias reflected in users' click histories. Future work could integrate political annotations directly into model prediction pipelines to assess whether recommender systems reinforce or exacerbate users' ideological preferences. In addition, NeMig enables the exploration of bias and diversity in multilingual and cross-lingual settings, as it provides German and English news content covering the same topical domain. Finally, designing recommender systems that balance predictive performance with diversity remains an open challenge. News recommenders should enable exposure to differing viewpoints rather than reinforcing existing beliefs. Having demonstrated how such biases can be quantified, one promising direction is to introduce objective functions that trades off accuracy for diversity, similarly to how the F_1-score balances precision and recall. Optimizing for such a target

function may be one pathway towards developing less biased news recommender systems. By introducing multilingual, politically annotated datasets, and establishing metrics for quantifying bias, this Chapter provides a foundation or developing more transparent, diverse, and responsible news recommender systems.

Having established a foundational understanding of the effects of bias in algorithmic news curation (**C2**), in the next Chapter we propose a solution to mitigate its detrimental consequences by introducing a framework for modular and customizable news recommendation. This framework enables the simultaneous adjustment of personalization and diversification levels in the generated recommendations (**C3**).

CHAPTER 7

MODULAR MULTI-ASPECT NEURAL NEWS RECOMMENDATION

Le seul véritable voyage ... ce ne serait pas d'aller vers de nouveaux paysages, mais d'avoir d'autres yeux, de voir l'univers avec les yeux d'un autre, de cent autres, de voir les cent univers que chacun d'eux voit ...
(The only true voyage of discovery ... would be not to visit new landscapes, but to possess other eyes, to see the universe through the eyes of another, of a hundred others, to see the hundred universes that each of them sees.)

MARCEL PROUST
«LA PRISONNIÈRE, À LA RECHERCHE DU TEMPS PERDU»

* The analyses of the previous Chapter reveal that modern news recommender systems face a dual challenge: they must provide accurate recommendations, while managing exposure diversity and mitigating content- and user-specific biases. At the same time, users' engagement with news is shaped by multiple aspects, such as topic, stance, or political orientation. Addressing these diverse requirements calls for recommendation frameworks that can flexibly integrate multiple dimensions of news characteristics and user preferences simultaneously, thereby helping to mitigate bias and improve exposure diversity.

Recent NNRs extend content-based recommendation (1) by aligning additional aspects (e.g., topic, sentiment) between candidate news and user history or (2) by diversifying recommendations w.r.t. these aspects. This customization is achieved by "hardcoding" additional constraints into the NNR's architecture and/or training objectives: any change in the desired recommendation behavior thus requires retraining the model with a modified objective (**C3**). This rigidity limits the widespread adoption of multi-aspect news recommenders.

In this Chapter, we introduce MANNeR, a modular framework for multi-aspect neural news recommendation that supports on-the-fly customization over individual aspects at inference time. With metric-based learning as its backbone, MANNeR learns aspect-specialized news encoders and then flexibly and linearly combines the resulting aspect-specific similarity scores into different ranking functions, alleviating the need for ranking function-specific retraining of the model. Extensive experimental results show that MANNeR consistently outperforms state-of-the-art NNRs on both standard content-based recommendation and single- and multi-aspect customization. Lastly, we validate that MANNeR's aspect-customization module is robust to language and domain transfer, making it applicable even in cross-lingual and low-resource settings.

*This Chapter is adapted from: **Andreea Iana**, Goran Glavaš, and Heiko Paulheim. 2024. Train Once, Use Flexibly: A Modular Framework for Multi-Aspect Neural News Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9555–9571, Miami, Florida, USA. Association for Computational Linguistics.

7.1 Introduction

Neural content-based recommenders, trained to infer users' preferences from their click history, represent the state of the art in news recommendation (Li and Wang, 2019; Wu et al., 2023). While previously consumed content clearly indicates users' preferences, *aspects* other than content alone, namely categorical features of the news such as topical category, sentiment, news outlet, or stance, contribute to their news consumption decisions. Accordingly, some NNRs leverage information on these aspects in addition to text content, be it (i) directly as model input (Wu et al., 2019a; Liu et al., 2020a) or (ii) indirectly, as auxiliary training tasks (Wu et al., 2019c, 2020c).

Increased personalization is often at odds with *diversity* (Pariser, 2011). NNRs optimized to maximize congruity to users' preferences tend to produce suggestions highly similar in content to previously consumed news (Liu et al., 2021b; Wu et al., 2020c; Sertkan and Neidhardt, 2023). Another strand of work thus focuses on increasing diversity of recommendations w.r.t. aspects other than content (e.g., sentiment). To this effect, prior work either (i) re-ranks content-based recommendations to decrease the aspectual similarity between them (Rao et al., 2013; Gharahighehi and Vens, 2023, *inter alia*), or (ii) trains the neural recommender by combining a content-based personalization objective with an aspect-based diversification objective (Wu et al., 2020c, 2022c; Shi et al., 2022; Choi et al., 2022).

Different users assign different importance to various news aspects (e.g., following developing events requires maximization of content-based overlap with the user's recent history; in another use-case, a user may prefer content-wise diversification of recommendations, but within the same topic of interest). Moreover, with personalization and diversification as mutually conflicting goals, users should be able to seamlessly define their own optimal trade-offs between the two. The existing body of work is ill-equipped for such multi-aspect customization, because each set of preferences – i.e., to personalize or diversify for each aspect – requires a different NNR to be trained from scratch. Put differently, forcing global assumptions on personalization and diversification preferences (i.e., same for all users) into the model design and training prevents customization at inference time.

Contributions. We propose a *modular* framework for *Multi-Aspect* Neural News Recommendation (MANNeR) to address this limitation. It leverages metric-based contrastive learning to induce a dedicated news encoder for each aspect, starting from a pretrained language model (PLM). This way, we obtain linearly-combinable aspect-specific similarity scores for pairs of news, allowing us to define ad-hoc at inference a custom ranking function for each user, reflecting their preferences across all aspects. MANNeR's modular design allows customization for any recommendation objective specified over (i) standard (i.e., content-based) personalization, (ii) aspect-based diversification, and (iii) aspect-based personalization. It also makes MANNeR easily extendable: to support personalization and diversification over a new aspect (e.g., news outlet), one only needs to train the aspect-specific news encoder for that aspect. Through extensive experiments with *topical categories* and *sentiment* as additional aspects next to content itself, we find that MANNeR outperforms state-of-the-art NNRs on standard content-based recommendation. Thanks to its module-specific outputs being *linearly composable* between objectives, we show – without training numerous models with different objectives – that depending on the recommendation goals, one can either (i) vastly increase aspect diversity (e.g., over topics and sentiment) of recommendations or (ii) improve aspect-based personalization, while retaining much of the content-based personalization performance. Finally, we demonstrate that MANNeR with a multilingual PLM is robust to the (cross-lingual) transfer of aspect-based encoders.

7.2 Related Work

Personalized Neural News Recommendation. Neural content-based models have become the main vehicle of personalized news recommendation, replacing traditional recommenders relying on manual feature engineering (Wu et al., 2023). Exploiting users' past behavior as NNR supervision leads to

recommendations that are content-wise closest to previously consumed news, in contrast to methods based on non-personalized criteria (Son et al., 2013; Chen et al., 2017; Ludmann, 2017). More recent NNRs seek to augment content-based personalization by considering other aspects, such as categories, sentiment, emotions (Sertkan and Neidhardt, 2022b), entities (Iana et al., 2024), outlets, or recency (Wu et al., 2023). These are incorporated in the NNR either as additional input to the news encoder (Wang et al., 2018b; Gao et al., 2018a; Wu et al., 2019a; Liu et al., 2020a; Sheu and Li, 2020; Lu et al., 2020; Qi et al., 2021b; Xun et al., 2021), or in the form of an auxiliary training objective for the news encoder (Wu et al., 2019c, 2020c; Qi et al., 2021c).

Diversification. Personalized NNR reduces exposure to news dissimilar from those consumed in the past. Recommending "more of the same" constrains access to diverse viewpoints and information (Freedman and Sears, 1965; Heitz et al., 2022) and leads to homogeneous news diets and "filter bubbles" (Pariser, 2011), in turn reinforcing users' initial stances (Li and Wang, 2019). Consequently, a significant body of work attempts to diversify recommendations, either by re-ranking them to increase some measure of diversity (e.g. intra-list distance (Zhang and Hurley, 2008)) or by resorting to multi-task training (Gabriel De Souza et al., 2019; Wu et al., 2020c; Shi et al., 2022; Wu et al., 2022c; Choi et al., 2022; Raza, 2023), coupling the primary content-based personalization objective with auxiliary objectives that force aspect-based diversification.

Current NNR Limitations. Critically, existing approaches, by "hardcoding" aspectual requirements (i.e., personalization or diversification for an aspect) into the NNR's architecture and/or training objectives, cannot be easily adjusted for varying recommendation goals. Since even minor changes in the recommendation objective require retraining the NNR, current models are generally limited to fixed single-aspect recommendation scenarios (e.g., content-based personalization with topical diversification), and ill-equipped for multi-aspect customization. In this work, we rethink personalized news recommendation and propose a novel, modular multi-aspect recommendation framework that allows for ad-hoc creation of recommendation functions over aspects at inference time. This enables fundamentally different recommendation: one that lets each user define their own custom recommendation function, choosing the amount of personalization or diversification for each aspect.

7.3 MANNeR: Multi-Aspect Neural News Recommendation

Personalized news recommendation produces for each candidate news n^c and user u with corresponding click history $\mathcal{H} = [n_1^u, n_2^u, ..., n_N^u]$, a relevance score $s(n^c, u)$ that quantifies the candidate's relevance for the user (see Section 3.1). In line with our previous definitions from Chapter 6, we define an *aspect* A_p as a categorical variable that encodes a news attribute (e.g. its category, stance, sentiment, provider), where each news n_i can belong only to one value of A_p (e.g. if A_p is the topic, then n_i may take exactly one value from {*politics*, *sports*, ...}). As discussed in Section 7.2, aspects are additional dimensions next to content over which to tailor recommendations, whether by (i) personalizing or (ii) diversifying over them. Following Chapter 6, we define *aspect-based personalization* as the level of homogeneity between a user's recommendations and clicked news w.r.t. the distribution of aspect A_p. In contrast, we define *aspect-based diversity* as the level of uniformity of aspect A_p's distribution among the news in the recommendation list.

We next introduce our proposed *modular framework* MANNeR, illustrated in Fig. 7.1. Starting from a PLM, during (1) training, we reshape the PLM's representation space via contrastive learning, independently for each aspect; this results in one specialized news encoder for each aspect; at (2) inference, we can, depending on the recommendation task, aggregate the resulting aspect-specific similarity scores to produce a final ranking function.

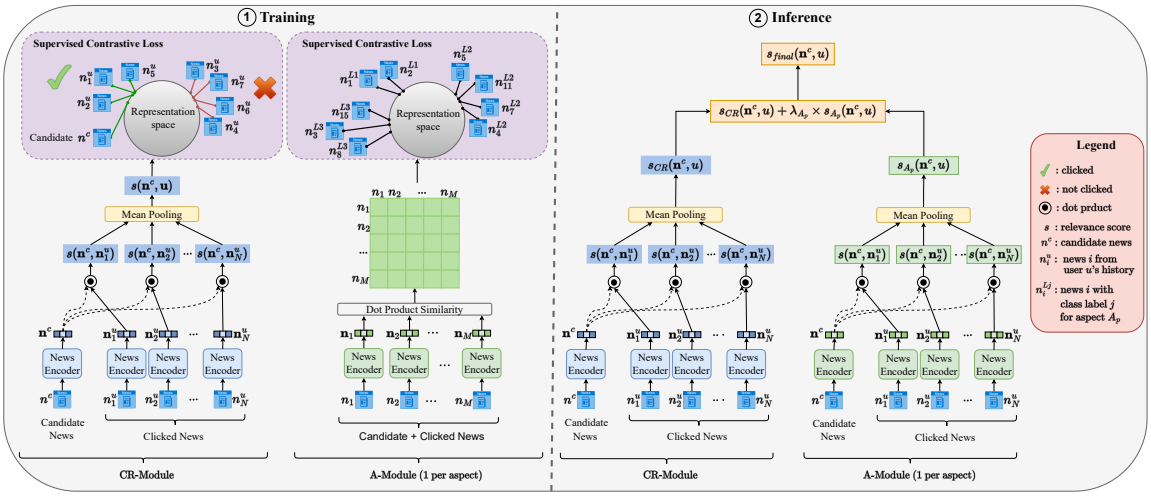


Figure 7.1: Illustration of the MANNeR framework. ① We train aspect-specialized NEs (i.e. `CR-Module` for content-based personalization, `A-Module` for aspect-based similarity) with metric-based contrastive learning. ② Inference: we linearly aggregate aspect-specific similarity scores between candidate and clicked news for final ranking.

7.3.1 News Encoder

We adopt a dual-component architecture for the news encoder coupling (i) a text and (ii) an entity encoder (Qi et al., 2021c,e). The former, a PLM, transforms the text input (i.e., concatenation of news title and abstract) into a text-based news embedding $\mathbf{n}_t$, given by the PLM's output `[CLS]` token representation. The latter learns an entity-level news embedding $\mathbf{n}_e$ by contextualizing pretrained embeddings of named entities (i.e., extracted from title and abstract) in a layer that combines multi-head self-attention (Vaswani et al., 2017) and additive attention (Bahdanau et al., 2015). The final news embedding $\mathbf{n}$ is the concatenation of $\mathbf{n}_t$ and $\mathbf{n}_e$.

7.3.2 Modular Training

MANNeR comprises two module types, each with a dedicated news encoder, responsible for content-based (`CR-Module`) and aspect-based (`A-Module`) recommendation relevance, respectively. We train both by minimizing the supervised contrastive loss (SCL, Equation 7.1) which aims to reshape the news encoder's representation space so that embeddings of same-class instances become mutually closer (cf. a distance/similarity metric) than instances of different classes (Khosla et al., 2020; Gunel et al., 2020b). To this end, we contrast the similarity score of a positive example (pair of same-class instances) against scores of corresponding negative examples (paired instances from different classes):

$$\mathcal{L} = -\sum_{i=1}^{N} \frac{1}{N_{y_i} - 1} \sum_{\substack{j \in [1,N] \\ i \neq j, y_i = y_j}} \log \frac{e^{(\mathbf{n}_i \cdot \mathbf{n}_j / \tau)}}{\sum_{\substack{k \in [1,N] \\ i \neq k}} e^{(\mathbf{n}_i \cdot \mathbf{n}_k / \tau)}} \tag{7.1}$$

with y_i as news n_i's label, N the batch size, N_{y_i} the number of batch instances with label y_i, and $\tau > 0$ the temperature hyperparameter controlling the extent of class separation. We use the dot product as the similarity metric for both module types.

CR-Module. Our `CR-Module` is a modification of the common content-based NNR architecture (Wu et al., 2023). Concretely, we encode both candidate and clicked news with a dedicated news encoder. However, following Chapter 4, we replace the widely used user encoder (i.e., early fusion of clicked news representations) with the simpler (and non-parameterized) mean-pooling of dot-product scores

between the candidate embedding $\mathbf{n}^c$ and clicked news embeddings $\mathbf{n}_i^u$: $s(\mathbf{n}^c, u) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{n}^c \cdot \mathbf{n}_i^u$ (i.e., late-fusion). We thus reduce the computational complexity of the standard approaches with elaborate parameterized user encoders. We then update the `CR-Module`'s encoder (i.e., fine-tune the PLM) by minimizing SCL, with clicked candidates as positive and non-clicked news as negative examples for the user. As there are many more non-clicked news, we resort to negative sampling (Wu et al., 2022b).

A-Module. Each `A-Module` trains a specialized news encoder for one aspect other than content. Via the metric-based objective, we reshape the PLM's representation space to group news according to aspect classes. Given a multi-class aspect, we first construct the training set from the union of all news in the dataset. Sets of news with the same aspect label form the positive samples for SCL; we obtain the corresponding negatives by pairing the same news from positive pairs with news from other aspect classes (e.g., for topical category as A_p, a news from *sports* is paired with the news from *politics* and/or *weather*). For each aspect, we independently fine-tune a separate copy of the same initial PLM. Note that the resulting aspect-specific news encoder encodes no information on user preferences: it only encodes the news similarity w.r.t. the aspect in question. Importantly, this implies that extending MANNeR to support a new aspect amounts to merely training an additional `A-Module` for that aspect.

7.3.3 Inference: Custom Ranking Functions

At inference time, the news encoders of the `CR-Module` and of each of the `A-Modules` are leveraged identically: we encode the candidate news as well as the user's clicked news with the respective news encoder, obtaining their module-specific embeddings $\mathbf{n}^c$ and $\mathbf{n}_i^u$ – their dot product $s = \mathbf{n}^c \cdot \mathbf{n}_i^u$ quantifies their similarity according to the module's aspect (or content for `CR-Module`'s news encoder). As different news encoders produce similarity scores of different magnitudes, we z-score normalize each module's scores per user. The final ranking score constitutes a *linear* aggregation of the content s_{CR} and aspect s_{A_p} similarity scores:

$$s_{final}(\mathbf{n}^c, \mathbf{u}) = s_{CR} + \sum_{A_p \in A} \lambda_{A_p} s_{A_p} \tag{7.2}$$

where λ_{A_p} is the scaling weight for the aspect score, and A the set of all aspects of interest. This linear composability of aspect-specific similarity scores allows not only generalization to multi-aspect recommendation objectives, but also different ad-hoc realizations of the ranking function that match custom recommendation goals: (i) with $\lambda_{A_p} = 0$, MANNeR performs standard content-based personalization, (ii) for $\lambda_{A_p} > 0$ it recommends based on both content- and aspect personalization, whereas (iii) for $\lambda_{A_p} < 0$ it simultaneously personalizes by content but diversifies for the aspect(s).

7.4 Experimental Setup

We compare MANNeR against state-of-the-art NNRs on a range of single- and multi-aspect recommendation tasks. We experiment with two aspects: *topical categories* (*ctg*) and *sentiment* (*snt*).

Baselines. We evaluate several NNRs trained on classification objectives. We follow Wu et al. (2021b) and replace the original news encoder of all baselines that do not use PLMs (instead, contextualizing word embeddings with convolution or self-attention layers) with the same PLM used in MANNeR.[1] We include two models optimized purely for content personalization: (1) NRMS (Wu et al., 2019e), and (2) MINER (Li et al., 2022). We further evaluate seven NNRs that inject aspect information. Thereof, five incorporate *topical categories*, i.e., (3) NAML (Wu et al., 2019a), (4) LSTUR (An et al.,

[1]The only exception is the final text embedding, where Wu et al. (2021b) pool tokens with an attention network.

2019), (5) MINS (Wang et al., 2022c), (6) CAUM (Qi et al., 2022b), (7) TANR (Wu et al., 2019c), and two the news *sentiment*: (8) SentiRec (Wu et al., 2020c), and (9) SentiDebias (Wu et al., 2022h).

Data. We carry out the evaluation on two prominent monolingual news recommendation benchmarks: MINDlarge (Wu et al., 2020e) with news in English and Adressa-1 week (Gulla et al., 2017) (denoted Adressa) with Norwegian news. For details on the dataset splits and corresponding statistics, we refer the reader to § 3.2.2. As Adressa contains no disambiguated named entities, we use only the news title as input to MANNeR' news encoder, while on MINDlarge we use all news features as input.

Regarding aspects, the topical category annotations are provided in both datasets. As no sentiment labels exist in neither MINDlarge nor Adressa, we use a multilingual XLM-RoBERTa$_{\text{Base}}$ model (Conneau et al., 2020) trained on tweets and fine-tuned for sentiment analysis (Barbieri et al., 2022b) to classify news into three classes: positive (pos), neutral, and negative (neg). We compute real-valued scores using the model's confidence scores s_i for class i, and the predicted sentiment class label $\hat{l}$ as:

$$s_{sent} = \begin{cases} (+1) \times s_{pos}, \text{ if } \hat{l}=pos \\ (-1) \times s_{neg}, \text{ if } \hat{l}=neg \\ (1 - s_{neutral}) \times (s_{pos} - s_{neg}), \text{ otherwise} \end{cases} \tag{7.3}$$

Evaluation Metrics. We report performance with AUC, MRR, NDCG@k ($k=\{5, 10\}$). Following Chapter 6, we measure aspect-based diversity of recommendations at position k as the normalized entropy of the distribution of aspect A_p's values in the recommendation list:

$$D_{A_p}@k = -\sum_{j \in A_p} \frac{p(j) \log p(j)}{\log(|A_p|)} \tag{7.4}$$

where $A_p \in \{ctg, snt\}$, and $|A_p|$ is the number of A_p classes. If aspect-based personalization is successful, aspect A_p's distribution in the recommendations should be similar to its distribution in the user history. We evaluate personalization with the generalized Jaccard similarity (Bonnici, 2020):

$$PS_{A_p}@k = \frac{\sum_{j=1}^{|A_p|} \min(\mathcal{R}_j, \mathcal{H}_j)}{\sum_{j=1}^{|A_p|} \max(\mathcal{R}_j, \mathcal{H}_j)}, \tag{7.5}$$

where $\mathcal{R}_j$ and $\mathcal{H}_j$ represent the probability of a news with class j of A_p to be contained in the recommendations list $\mathcal{R}$, and, respectively, in the user history $\mathcal{H}$. As all metrics are bounded to $[0, 1]$, we measure the trade-off between content-based personalization (NDCG@k) and either aspect-based diversity $D_{A_p}@k$ or aspect-based personalization $PS_{A_p}@k$ with the harmonic mean. We denote this $\text{T}_{\text{A}_\text{p}}@k$ for single-aspect. For multi-aspect evaluation, i.e., when ranking for content-personalization by diversifying simultaneously over topics and sentiment, we adopt as evaluation metric the harmonic mean between NDCG@k, $\text{D}_{\text{ctg}}@k$ (topical category), and $\text{D}_{\text{snt}}@k$ (sentiment), denoted $\text{T}_{\text{all}}@k$.

Training Details. We use RoBERTa$_{\text{Base}}$ (Liu et al., 2019b) and NB-BERT$_{\text{Base}}$ (Kummervold et al., 2021; Nielsen, 2023) in experiments on MINDlarge and Adressa, respectively. We set the maximum history length to 50. We tune the main hyperparameters of all NNRs. We train all models with mixed precision, the Adam optimizer (Kingma and Ba, 2014), the learning rate of 1e-5 on MINDlarge, 1e-6 on Adressa, and 1e-6 for the sentiment`A-Module` on both datasets. In `A-Module` training, we sample 20 instances per class,[2] while in `CR-Module` training we sample four negatives per positive example. We find the optimal temperature of 0.36 on MINDlarge, and 0.14 on Adressa, for the `CR-Module`, and of 0.9 for all `A-Modules` on both datasets. We train all baselines and the `CR-Module` for 5 epochs on MINDlarge and 20 epochs on Adressa, with a batch size of 8. We train each `A-Module` for 100 epochs, with the batch size of 60 for sentiment and 360 for topics. We repeat runs five times with different

[2]For M class instances, we obtain $\frac{M^2-M}{2}$ positive pairs for that class for SCL.

	MINDlarge				Adressa			
Model	AUC	MRR	NDCG@5	NDCG@10	AUC	MRR	NDCG@5	NDCG@10
NRMS-PLM	$63.02_{\pm 1.46}$	$35.47_{\pm 0.60}$	$33.43_{\pm 0.71}$	$39.87_{\pm 0.62}$	$72.32_{\pm 3.28}$	$43.02_{\pm 2.65}$	$44.34_{\pm 2.83}$	$51.29_{\pm 2.31}$
MINER	$63.10_{\pm 1.17}$	$35.48_{\pm 1.11}$	$33.67_{\pm 1.14}$	$40.00_{\pm 1.04}$	$70.11_{\pm 4.95}$	$37.33_{\pm 4.09}$	$38.51_{\pm 5.08}$	$46.26_{\pm 4.11}$
NAML-PLM	$60.64_{\pm 3.36}$	$37.58_{\pm 0.41}$	$35.91_{\pm 0.41}$	$42.23_{\pm 0.40}$	$50.00_{\pm 0.00}$	$44.97_{\pm 5.01}$	$47.24_{\pm 5.53}$	$52.53_{\pm 4.15}$
LSTUR-PLM	$54.59_{\pm 2.97}$	$33.35_{\pm 1.52}$	$31.71_{\pm 1.78}$	$38.31_{\pm 1.70}$	$65.01_{\pm 7.16}$	$43.14_{\pm 1.70}$	$44.82_{\pm 2.57}$	$51.19_{\pm 2.05}$
MINS-PLM	$60.70_{\pm 2.77}$	$36.26_{\pm 0.36}$	$34.47_{\pm 0.45}$	$40.82_{\pm 0.32}$	$74.25_{\pm 3.16}$	$44.18_{\pm 2.94}$	$47.30_{\pm 3.26}$	$52.97_{\pm 3.37}$
$\text{CAUM}_{\text{no entities}}$-PLM	$66.16_{\pm 2.97}$	$36.30_{\pm 1.96}$	$34.56_{\pm 1.98}$	$40.99_{\pm 1.91}$	$76.51_{\pm 1.19}$	$43.62_{\pm 1.31}$	$46.91_{\pm 1.28}$	$51.99_{\pm 1.23}$
CAUM-PLM	$66.35_{\pm 3.06}$	$36.22_{\pm 1.20}$	$34.34_{\pm 1.29}$	$40.81_{\pm 1.32}$	–	–	–	–
TANR-PLM	$63.34_{\pm 1.06}$	$36.97_{\pm 0.98}$	$35.17_{\pm 1.04}$	$41.62_{\pm 0.93}$	$50.00_{\pm 0.00}$	$43.82_{\pm 0.97}$	$45.64_{\pm 1.31}$	$51.44_{\pm 0.60}$
SentiRec-PLM	$62.18_{\pm 0.70}$	$35.75_{\pm 0.42}$	$33.88_{\pm 0.43}$	$40.46_{\pm 0.40}$	$67.59_{\pm 2.70}$	$33.06_{\pm 2.36}$	$32.85_{\pm 3.76}$	$40.78_{\pm 2.39}$
SentiDebias-PLM	$55.00_{\pm 2.51}$	$27.76_{\pm 1.86}$	$25.49_{\pm 1.94}$	$32.21_{\pm 1.95}$	$67.42_{\pm 2.38}$	$35.68_{\pm 3.40}$	$36.45_{\pm 4.21}$	$44.21_{\pm 2.93}$
MANNeR (`CR-Module`)	$\mathbf{69.74}_{\pm \mathbf{0.85}}$	$\mathbf{38.64}_{\pm \mathbf{0.62}}$	$\mathbf{37.01}_{\pm \mathbf{0.61}}$	$\mathbf{43.18}_{\pm \mathbf{0.57}}$	$\mathbf{79.37}_{\pm \mathbf{1.69}}$	$\mathbf{47.03}_{\pm \mathbf{2.39}}$	$\mathbf{48.88}_{\pm \mathbf{2.78}}$	$\mathbf{54.30}_{\pm \mathbf{2.48}}$
Improvement (%)	+ 5.41	+ 2.82	+ 3.06	+ 2.25	+ 3.73	+ 4.58	+ 3.34	+ 2.51

Table 7.1: Content-based recommendation performance. We average results across five runs, and report the relative improvement over the best baseline. The best results per column are highlighted in bold, the second best underlined.

seeds and report averages and standard deviations for all metrics. We refer to Appendices E.1.1 and E.1.2 for further details about model sizes and hyperparameters.

7.5 Results and Discussion

We first discuss MANNeR's content personalization performance. We then analyze its capability for single- and multi-aspect (i) diversification and (ii) personalization. In the aspect customization setups, we treat MANNeR's `CR-Module` as a baseline. Lastly, we evaluate its ability to re-use pretrained aspect-specific modules in cross-lingual transfer.

7.5.1 Content Personalization

Table 7.1 summarizes the results on content personalization. As the task does not require aspect-based customization, we evaluate the MANNeR variant that uses only its CR-Module at inference time (i.e., $\lambda = 0$). MANNeR consistently outperforms all state-of-the-art NNRs in terms of both classification and ranking metrics on both datasets. Given that MANNeR's `CR-Module` derives the user embedding by merely averaging clicked news embeddings, these results question the need for complex parameterized user encoders present in all baselines, in line with the findings of Chapter 4.

We ablate the `CR-Module`'s content personalization performance for (i) different inputs to the news encoder and (ii) alternative architecture designs and training objectives. We find that all groups of features (e.g., abstract, named entities) contribute to the overall performance (cf. Fig. E.1a). Moreover, we confirm the findings of Chapter 4 that (i) late fusion outperforms a parameterized user encoder (i.e., early fusion), and that (ii) SCL better separates classes than cross-entropy loss, in line with other similarity-oriented NLP tasks (Reimers and Gurevych, 2019).

7.5.2 Single-Aspect Customization

Diversification. Table 7.2 summarizes the results on aspect diversification tasks. Most baselines (including MANNeR's `CR-Module` without aspect diversification) obtain similar diversification scores (D_{ctg} and D_{snt}). The sentiment-aware SentiRec-PLM, with an explicit auxiliary sentiment diversification objective, yields the highest sentiment diversity on Adressa; this comes at the expense of content personalization quality (lowest NDCG). On MINDlarge, the sentiment-specific SentiDebias-PLM achieves the highest sentiment diversity, but also exhibits lower content personalization performance. Overall, these results point to a trade-off between content personalization and aspectual diversity: models with higher D_{A_p} tend to have a lower NDCG.

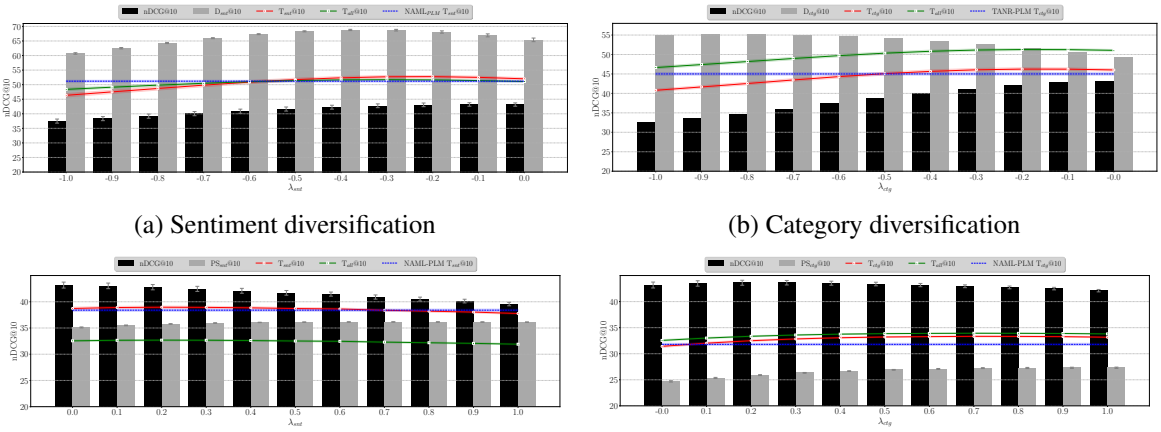
Model	NDCG@10	D_{ctg}@10	T_{ctg}@10	D_{snt}@10	T_{snt}@10	T_{all}@10
	MINDlarge					
NRMS-PLM	39.87$_{\pm 0.62}$	49.98$_{\pm 1.06}$	44.35$_{\pm 0.3}$	66.41$_{\pm 0.49}$	49.83$_{\pm 0.46}$	49.87$_{\pm 0.30}$
MINER	40.00$_{\pm 1.04}$	49.45$_{\pm 1.22}$	44.21$_{\pm 0.44}$	65.68$_{\pm 0.89}$	49.71$_{\pm 0.98}$	49.61$_{\pm 0.52}$
NAML-PLM	42.23$_{\pm 0.40}$	47.33$_{\pm 0.30}$	44.63$_{\pm 0.34}$	65.14$_{\pm 0.38}$	51.24$_{\pm 0.34}$	49.86$_{\pm 0.30}$
LSTUR-PLM	38.31$_{\pm 1.70}$	50.04$_{\pm 1.17}$	43.36$_{\pm 0.69}$	65.62$_{\pm 0.34}$	48.35$_{\pm 1.27}$	48.88$_{\pm 0.53}$
MINS-PLM	40.82$_{\pm 0.32}$	48.87$_{\pm 0.96}$	44.48$_{\pm 0.22}$	66.20$_{\pm 0.97}$	50.50$_{\pm 0.13}$	49.94$_{\pm 0.37}$
$\text{CAUM}_{\text{no entities}}$-PLM	40.99$_{\pm 1.91}$	47.45$_{\pm 1.04}$	43.94$_{\pm 0.88}$	65.84$_{\pm 1.16}$	50.49$_{\pm 1.31}$	49.41$_{\pm 0.64}$
CAUM-PLM	40.81$_{\pm 1.32}$	47.85$_{\pm 0.89}$	44.04$_{\pm 1.03}$	66.52$_{\pm 0.53}$	50.57$_{\pm 1.03}$	49.63$_{\pm 0.91}$
TANR-PLM	41.62$_{\pm 0.93}$	48.92$_{\pm 0.92}$	44.96$_{\pm 0.28}$	66.05$_{\pm 0.85}$	51.06$_{\pm 0.66}$	50.32$_{\pm 0.34}$
SentiRec-PLM	40.46$_{\pm 0.40}$	49.43$_{\pm 0.39}$	44.50$_{\pm 0.13}$	66.99$_{\pm 0.60}$	50.45$_{\pm 0.38}$	50.10$_{\pm 0.16}$
SentiDebias-PLM	32.21$_{\pm 1.95}$	**52.03$_{\pm 2.17}$**	39.72$_{\pm 1.05}$	68.59$_{\pm 1.25}$	43.80$_{\pm 1.81}$	46.20$_{\pm 0.98}$
MANNeR (`CR-Module`)	**43.18$_{\pm 0.57}$**	49.29$_{\pm 0.29}$	46.03$_{\pm 0.29}$	65.39$_{\pm 0.62}$	52.01$_{\pm 0.40}$	51.07$_{\pm 0.22}$
MANNeR ($\lambda_{ctg} = -0.2$, $\lambda_{snt} = 0$)	41.97$_{\pm 0.61}$	51.50$_{\pm 0.32}$	**46.24$_{\pm 0.35}$**	65.56$_{\pm 0.60}$	51.17$_{\pm 0.44}$	51.28$_{\pm 0.26}$
MANNeR ($\lambda_{ctg} = 0$, $\lambda_{snt} = -0.3$)	42.75$_{\pm 0.65}$	49.83$_{\pm 0.23}$	46.02$_{\pm 0.36}$	**68.71$_{\pm 0.28}$**	**52.71$_{\pm 0.43}$**	**51.71$_{\pm 0.26}$**
	Adressa					
NRMS-PLM	51.29$_{\pm 2.31}$	31.79$_{\pm 1.04}$	39.21$_{\pm 0.46}$	61.54$_{\pm 0.53}$	55.92$_{\pm 1.21}$	44.60$_{\pm 0.45}$
MINER	46.26$_{\pm 4.11}$	31.09$_{\pm 0.65}$	37.14$_{\pm 1.65}$	60.88$_{\pm 0.47}$	52.51$_{\pm 2.77}$	42.68$_{\pm 1.52}$
NAML-PLM	52.53$_{\pm 4.15}$	30.58$_{\pm 2.35}$	38.55$_{\pm 2.13}$	61.64$_{\pm 0.58}$	56.66$_{\pm 2.63}$	44.04$_{\pm 1.95}$
LSTUR-PLM	51.19$_{\pm 2.05}$	29.93$_{\pm 4.63}$	37.65$_{\pm 4.20}$	61.37$_{\pm 0.54}$	55.80$_{\pm 1.22}$	43.15$_{\pm 3.76}$
MINS-PLM	52.97$_{\pm 3.37}$	33.61$_{\pm 1.70}$	41.04$_{\pm 1.02}$	61.79$_{\pm 0.62}$	56.99$_{\pm 1.76}$	46.21$_{\pm 0.90}$
$\text{CAUM}_{\text{no entities}}$-PLM	51.99$_{\pm 1.23}$	34.39$_{\pm 0.27}$	**41.39$_{\pm 0.36}$**	62.06$_{\pm 0.47}$	56.57$_{\pm 0.72}$	**46.56$_{\pm 0.31}$**
TANR-PLM	51.44$_{\pm 0.60}$	32.89$_{\pm 1.69}$	40.09$_{\pm 1.11}$	61.80$_{\pm 0.65}$	56.14$_{\pm 0.20}$	45.41$_{\pm 1.05}$
SentiRec-PLM	40.78$_{\pm 2.39}$	**35.61$_{\pm 0.60}$**	37.99$_{\pm 1.08}$	**68.54$_{\pm 0.18}$**	51.10$_{\pm 1.91}$	44.61$_{\pm 1.00}$
SentiDebias-PLM	44.21$_{\pm 2.93}$	32.29$_{\pm 1.02}$	37.28$_{\pm 1.21}$	61.24$_{\pm 0.24}$	51.31$_{\pm 2.03}$	42.87$_{\pm 1.07}$
MANNeR (`CR-Module`)	**54.30$_{\pm 2.48}$**	31.74$_{\pm 0.15}$	40.05$_{\pm 0.70}$	61.37$_{\pm 0.34}$	57.60$_{\pm 1.46}$	45.29$_{\pm 0.63}$
MANNeR ($\lambda_{ctg} = -0.3$, $\lambda_{snt} = 0$)	50.93$_{\pm 2.53}$	34.06$_{\pm 0.25}$	40.80$_{\pm 0.77}$	61.87$_{\pm 0.34}$	55.85$_{\pm 1.58}$	46.02$_{\pm 0.70}$
MANNeR ($\lambda_{ctg} = 0$, $\lambda_{snt} = -0.2$)	53.80$_{\pm 2.51}$	32.39$_{\pm 0.17}$	40.42$_{\pm 0.68}$	63.04$_{\pm 0.29}$	**58.03$_{\pm 1.35}$**	45.91$_{\pm 0.62}$

Table 7.2: Single-aspect diversification. For MANNeR, we list the best results (cf. T_{A_p}) of single-aspect diversification as λ_{A_p}. The best results per dataset and per column are highlighted in bold, the second best underlined.

Unlike all other models, MANNeR can trade content personalization for diversity (and vice-versa) with different values of the aspect coefficients λ_{A_p}. Figs. 7.2a and 7.2b illustrate its performance in single-aspect sentiment and category diversification tasks for different values of λ_{snt}, and λ_{ctg}, respectively, on MINDlarge. The steady drop in NDCG together with the steady increase in D_{A_p} indeed indicate the existence of a trade-off between content personalization and aspect diversification. For topical categories we observe a steeper decline in content personalization quality with improved diversification than for sentiment. Sentiment diversity reaches peak performance for $\lambda_{snt} = -0.4$, whereas category diversity continues to increase up to $\lambda_{ctg} = -0.9$. Intuitively, content-based recommendation is more aligned with the topical than with the sentiment consistency of recommendations. The best trade-off (i.e., maximal performance w.r.t. T_{A_p}@10) is achieved for $\lambda_{snt} = -0.3$ for sentiment, and $\lambda_{ctg} = -0.2$ for topics.[3] We attribute these effects to the representation spaces of the `A-Modules`. Figs. 7.3a and 7.3b shows the 2-dimensional t-SNE visualizations (Van der Maaten and Hinton, 2008) of the news embeddings produced with category-specialized, and respectively, sentiment-specialized news encoders trained on MINDlarge. The results confirm that the encoder's latent representation space was reshaped to group same-class instances. The separation of classes, however, is less prominent for the representation spaces of the encoders trained on Adressa (cf. Figs. 7.3c and 7.3d , e.g., the effect is stronger on the category-shaped embedding space).[4]

[3] We report analogous results on Adressa in Figs. E.2a and E.2b.

[4] We believe that this is because Adressa has 10 times fewer news than MINDlarge, with over half of the topical categories in Adressa being represented with fewer than 100 examples.



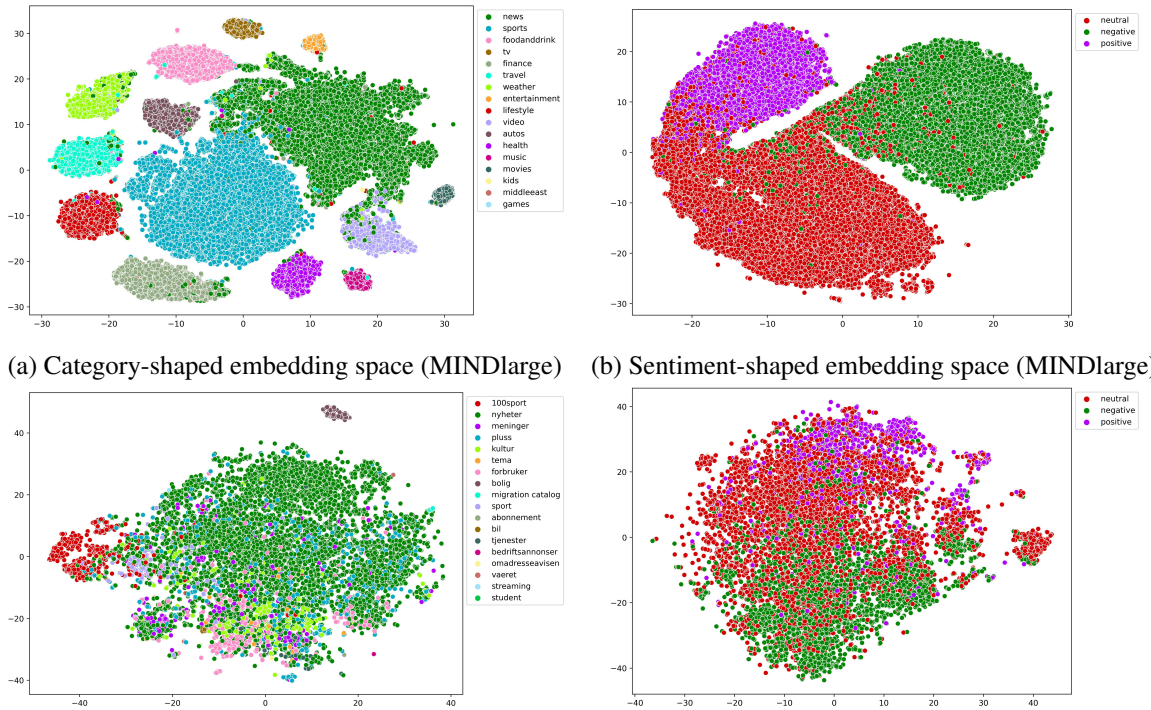
(a) Sentiment diversification

(b) Category diversification

(c) Sentiment personalization

(d) Category personalization

Figure 7.2: Results of single-aspect customization for MANNeR and the best baseline on MINDlarge.



(a) Category-shaped embedding space (MINDlarge)

(b) Sentiment-shaped embedding space (MINDlarge)

(c) Category-shaped embedding space (Adressa)

(d) Sentiment-shaped embedding space (Adressa)

Figure 7.3: t-SNE plots of the news embeddings in the test sets of MINDlarge and Adressa.

Personalization. Table 7.3 displays the results on aspect personalization tasks. TANR, trained with an auxiliary topic classification task, underperforms NAML, which uses topical categories as news encoder input features, in category personalization on both datasets. MANNeR's `CR-Module` alone (i.e., without any aspect customization) yields competitive category personalization performance. We believe that this is because (i) the `CR-Module` is best in content personalization and (ii) category personalization is well-aligned with content personalization (i.e., news with similar content tend to belong to the same category). Fig. 7.2d explores the trade-off between content and category personalization, for positive values of λ_{ctg} on MINDlarge. The best topical category personalization (PS_{ctg}), obtained for $\lambda_{ctg} > 0.7$, comes at the small expense of content personalization: too much weight on the category similarity of news dilutes the impact of content relevance. Increased sentiment personalization (cf. Fig. 7.2c), however, is much more detrimental to content personalization.

Model	NDCG@10	PS_{ctg}@10	T_{ctg}@10	PS_{snt}@10	T_{snt}@10	T_{all}@10
	MINDlarge					
NRMS-PLM	$39.87_{\pm 0.62}$	$23.94_{\pm 0.16}$	$29.92_{\pm 0.27}$	$35.05_{\pm 0.15}$	$37.31_{\pm 0.25}$	$31.45_{\pm 0.20}$
MINER	$40.00_{\pm 1.04}$	$23.94_{\pm 0.38}$	$29.95_{\pm 0.50}$	$34.98_{\pm 0.21}$	$37.32_{\pm 0.36}$	$31.46_{\pm 0.35}$
NAML-PLM	$42.23_{\pm 0.40}$	$25.47_{\pm 0.21}$	$31.77_{\pm 0.18}$	$35.14_{\pm 0.15}$	$38.35_{\pm 0.16}$	$32.82_{\pm 0.16}$
LSTUR-PLM	$38.31_{\pm 1.70}$	$24.01_{\pm 0.99}$	$29.51_{\pm 1.21}$	$34.81_{\pm 0.33}$	$36.46_{\pm 0.90}$	$31.08_{\pm 0.98}$
MINS-PLM	$40.82_{\pm 0.32}$	$25.00_{\pm 0.36}$	$31.01_{\pm 0.35}$	$34.81_{\pm 0.19}$	$37.58_{\pm 0.17}$	$32.18_{\pm 0.28}$
$CAUM_{\text{no entities}}$-PLM	$40.99_{\pm 1.91}$	$24.84_{\pm 0.60}$	$30.93_{\pm 0.98}$	$35.03_{\pm 0.19}$	$37.76_{\pm 0.86}$	$32.18_{\pm 0.72}$
CAUM-PLM	$40.81_{\pm 1.32}$	$25.07_{\pm 0.27}$	$31.06_{\pm 0.45}$	$34.96_{\pm 0.13}$	$37.65_{\pm 0.61}$	$32.26_{\pm 0.35}$
TANR-PLM	$41.62_{\pm 0.93}$	$25.19_{\pm 0.49}$	$31.38_{\pm 0.60}$	$35.02_{\pm 0.17}$	$38.03_{\pm 0.44}$	$32.51_{\pm 0.47}$
SentiRec-PLM	$40.46_{\pm 0.40}$	$24.21_{\pm 0.25}$	$30.30_{\pm 0.28}$	$34.59_{\pm 0.04}$	$37.29_{\pm 0.17}$	$31.60_{\pm 0.20}$
SentiDebias-PLM	$32.21_{\pm 1.95}$	$20.76_{\pm 1.28}$	$25.24_{\pm 1.48}$	$34.11_{\pm 0.31}$	$33.11_{\pm 1.17}$	$27.62_{\pm 1.25}$
MANNeR (`CR-Module`)	$\mathbf{43.18_{\pm 0.57}}$	$24.69_{\pm 0.13}$	$31.41_{\pm 0.21}$	$35.11_{\pm 0.09}$	$38.72_{\pm 0.24}$	$32.55_{\pm 0.15}$
MANNeR ($\lambda_{ctg} = 0.7, \lambda_{snt} = 0$)	$42.94_{\pm 0.27}$	$\mathbf{27.24_{\pm 0.09}}$	$\mathbf{33.33_{\pm 0.12}}$	$35.18_{\pm 0.05}$	$38.68_{\pm 0.09}$	$\mathbf{33.93_{\pm 0.07}}$
MANNeR ($\lambda_{ctg} = 0, \lambda_{snt} = 0.2$)	$42.78_{\pm 0.50}$	$24.72_{\pm 0.12}$	$31.34_{\pm 0.20}$	$\mathbf{35.77_{\pm 0.06}}$	$\mathbf{38.96_{\pm 0.21}}$	$32.69_{\pm 0.15}$
	Adressa					
NRMS-PLM	$51.29_{\pm 2.31}$	$34.31_{\pm 0.41}$	$41.11_{\pm 0.96}$	$41.85_{\pm 0.08}$	$46.07_{\pm 0.97}$	$41.35_{\pm 0.67}$
MINER	$46.26_{\pm 4.11}$	$34.42_{\pm 0.24}$	$39.41_{\pm 1.46}$	$41.95_{\pm 0.02}$	$43.93_{\pm 1.80}$	$40.22_{\pm 1.00}$
NAML-PLM	$52.53_{\pm 4.15}$	$36.06_{\pm 0.82}$	$42.71_{\pm 1.68}$	$41.82_{\pm 0.07}$	$46.51_{\pm 1.67}$	$42.40_{\pm 1.12}$
LSTUR-PLM	$51.19_{\pm 2.05}$	$35.11_{\pm 2.06}$	$41.59_{\pm 1.02}$	$41.83_{\pm 0.11}$	$46.03_{\pm 0.82}$	$41.66_{\pm 0.72}$
MINS-PLM	$52.97_{\pm 3.37}$	$33.86_{\pm 0.70}$	$41.29_{\pm 1.40}$	$41.80_{\pm 0.11}$	$46.69_{\pm 1.35}$	$41.45_{\pm 0.97}$
$CAUM_{\text{no entities}}$-PLM	$51.99_{\pm 1.23}$	$33.53_{\pm 0.19}$	$40.76_{\pm 0.36}$	$41.75_{\pm 0.08}$	$46.30_{\pm 0.51}$	$41.08_{\pm 0.25}$
TANR-PLM	$51.44_{\pm 0.60}$	$34.02_{\pm 0.50}$	$40.96_{\pm 0.54}$	$41.82_{\pm 0.08}$	$46.13_{\pm 0.28}$	$41.24_{\pm 0.38}$
SentiRec-PLM	$40.78_{\pm 2.39}$	$32.42_{\pm 0.30}$	$36.09_{\pm 0.97}$	$39.32_{\pm 0.11}$	$40.01_{\pm 1.17}$	$37.10_{\pm 0.69}$
SentiDebias-PLM	$44.21_{\pm 2.93}$	$34.11_{\pm 0.58}$	$38.48_{\pm 1.34}$	$41.76_{\pm 0.06}$	$42.92_{\pm 1.41}$	$39.51_{\pm 0.96}$
MANNeR (`CR-Module`)	$\mathbf{54.30_{\pm 2.48}}$	$34.50_{\pm 0.07}$	$42.18_{\pm 0.80}$	$41.99_{\pm 0.06}$	$47.34_{\pm 0.93}$	$42.11_{\pm 0.53}$
MANNeR ($\lambda_{ctg} = 0.4, \lambda_{snt} = 0$)	$53.62_{\pm 1.93}$	$\mathbf{36.25_{\pm 0.09}}$	$\mathbf{43.24_{\pm 0.66}}$	$42.12_{\pm 0.07}$	$47.17_{\pm 0.74}$	$\mathbf{42.86_{\pm 0.43}}$
MANNeR ($\lambda_{ctg} = 0, \lambda_{snt} = 0.1$)	$54.07_{\pm 2.37}$	$34.68_{\pm 0.06}$	$42.24_{\pm 0.76}$	$\mathbf{42.22_{\pm 0.07}}$	$\mathbf{47.40_{\pm 0.90}}$	$42.23_{\pm 0.51}$

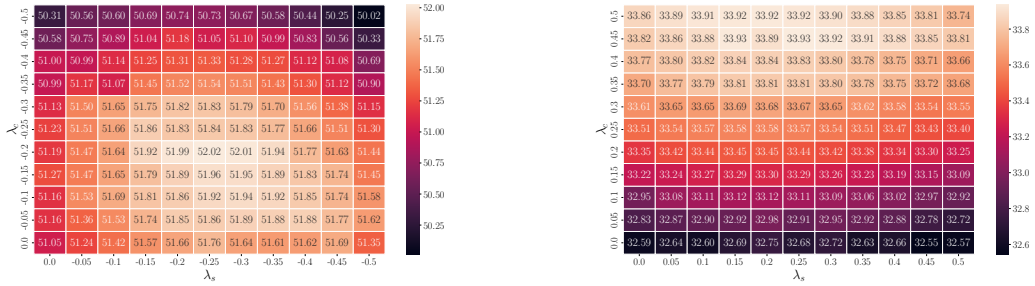
Table 7.3: Single-aspect personalization. For MANNeR, we list the best results (cf. T_{A_p}) of single-aspect diversification as λ_{A_p}). The best results per dataset and per column are highlighted in bold, the second best underlined.

Intuitively, users do not choose articles based on sentiment. Tailoring recommendations according to the sentiment of previously clicked news thus leads to more content-irrelevant suggestions.

7.5.3 Multi-Aspect Customization

We further explore the trade-off between content personalization and multi-aspect diversification, i.e. diversifying over both topical categories and sentiments. We achieve the highest T_{all} for $\lambda_{ctg} = -0.2$ and $\lambda_{snt} = -0.25$ on MINDlarge (cf. Fig. 7.4a). In line with single-aspect diversification results, we observe that improving diversity in terms of topical categories rather than sentiments has a more negative effect on content personalization quality, i.e. steeper decline in T_{all}. These results confirm that MANNeR can generalize to diversify for multiple aspects at once by weighting individual aspect relevance scores less than in the single-aspect task. Weighting several aspects higher simultaneously acts as a double discounting for content personalization, diluting content relevance disproportionately. Similarly, for multi-aspect personalization, we achieve the best multi-aspect trade-off on MINDlarge (cf. Fig. 7.4b) for $\lambda_{ctg} = 0.45$ and $\lambda_{snt} = 0.25$. Stronger enforcing of alignment of candidate news with the user's history is needed for topical categories than for sentiment (i.e., $\lambda_{ctg} > \lambda_{snt}$). This is because sentiment exhibits low variance within categories (e.g., *politics* news are mostly negative) and enforcing categorical personalization partly also achieves sentiment personalization.[5]

[5]We refer to Fig. E.3 for analogous results on Adressa.



(a) Multi-aspect diversification on MINDlarge

(b) Multi-aspect personalization on MINDlarge

Figure 7.4: Results of multi-aspect customization for MANNeR on MINDlarge.

7.5.4 Cross-Lingual Transfer

We next analyze the transferability of MANNeR across datasets and languages in single-aspect customization experiments.[6] Concretely, we train the `CR-Module` and `A-Modules` on both MINDlarge (i.e., in English) and Adressa (i.e., in Norwegian), respectively. At inference, we evaluate all combinations of pretrained `CR-Module` and `A-Modules` on the test set of MINDlarge. We now use a multilingual DistilBERT$_{\text{Base}}$ (Sanh et al., 2019) as MANNeR's news encoder to enable cross-lingual transfer (`XLT`). Fig. 7.5 summarizes the `XLT` results for single-aspect diversification and personalization, respectively, on the target-language dataset MIND.[7] As expected, MANNeR trained fully on Adressa suffers a large drop in content personalization performance, compared to the counterpart trained on MINDlarge. In contrast, transferring only the `A-Module`, i.e., training the `CR-Module` on MINDlarge and the `A-Module` on Adressa, yields performance comparable to that of complete in-language training (i.e., both `CR-Module` and `A-Module` trained on MINDlarge). This is particularly the case for the sentiment `A-Module`, since the sentiment labels between the datasets are more aligned than those for topical categories. These results indicate that the plug-and-play of `A-Modules` enables zero-shot `XLT`, as modules trained on the much smaller Norwegian Adressa transfer well to the large English MINDlarge. This suggests that, coupled with multilingual PLMs, MANNeR can be used for effective news recommendation in lower-resource languages, where training data and aspectual labels are scarce. Furthermore, the results demonstrate that the `A-Modules` could be trained on general-purpose classification datasets (e.g. topic or sentiment classification datasets), alleviating the need for aspect-specific annotation of news stories.

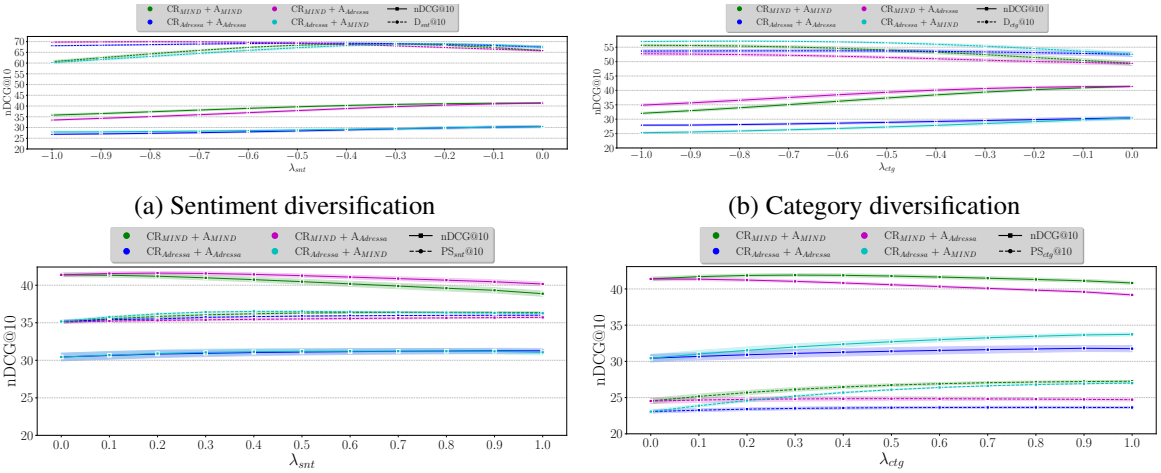
7.5.5 Computational Complexity

While A-Modules add extra parameters, their average training time is two orders of magnitude faster than that of the CR-Module.[8] This is a *one-time* increase in training time: the resulting modules can then be arbitrarily combined for any recommendation goal without additional training. In contrast, all other NNRs require re-training or fine-tuning if the recommendation objective changes as the model weights have to be adjusted each time. This translates into much higher computational costs in practice. We emphasize that MANNeR also has a much lower inference latency (see Appendix E.2.5) due to the (i) CR-Module's lean architecture without a parameterized user encoder, and (ii) ability to parallelize loading and deploying different modules, for which only the final score has to be combined. Overall, considering both training and inference, MANNeR is more efficient and flexible in a realistic setup with differing recommendation goals that may vary by user or for the same user over time.

[6] We evaluate only the title-based version of MANNeR, as the full version cannot be trained on Adressa.

[7] Fig. E.4 provide similar results for single-aspect customization on Adressa, as target-language datasets.

[8] On MINDlarge (Adressa), the A-Module for topical category trains 277 (51) times faster and that for sentiment 204 (53) times faster on average per epoch than the CR-Module.



(a) Sentiment diversification

(b) Category diversification

(c) Sentiment personalization

(d) Category personalization

Figure 7.5: XLT in single-aspect diversification and personalization, with modules trained on different (combinations of) source-language datasets and evaluated on the target-language dataset MINDlarge. The line style indicates the metric, the color the source-language datasets used in training.

7.6 Conclusion

We proposed MANNeR, a modular framework for multi-aspect neural news recommendation. It learns aspect-specialized NEs with supervised contrastive learning, and linearly combines the corresponding aspect-specific similarity scores for final ranking. Its modular design allows defining ad-hoc multi-aspect ranking functions at inference. Our experiments show that MANNeR consistently outperforms state-of-the-art NNRs on both (i) standard content-based recommendation, and on single- and multi-aspect (ii) diversification and (iii) personalization of recommendations. Moreover, we can identify on-the-fly optimal trade-offs between content-based recommendation performance and aspect-based customization. Equipped with a multilingual PLM, MANNeR can successfully cross-lingually transfer aspect-specific news encoders. This supports use cases where aspect-specific labels (e.g., sentiment) are not available for news in the target languages of interest. We hope that our work stimulates more research towards modular, easily extendable, and reusable news recommenders.

In this Chapter, we have presented a solution for adapting neural news recommenders to efficiently support multi-aspect customization beyond pure relevance (**C3**). In the following Chapters, we move to our last challenge which deals with the linguistic inclusivity in the news recommendation data and modeling approaches (**C4**).

Part IV

MULTILINGUALITY

CHAPTER 8

CROSS-LINGUAL NEWS RECOMMENDATION

The diversity of languages is not a diversity of signs and sounds but a diversity of views of the world.

WILHELM VON HUMBOLDT

[*] Language constitutes a fundamental medium of human interaction, reflecting the global diversity of cultures, perspectives, and knowledge. Arguably, language shapes not only how individuals communicate with one another, but also how they engage with information. In the context of news recommendation, language is intrinsically tied to content consumption: it influences which news users access, in which language they prefer to consume specific topics, and how information is interpreted (see § 1.1.2). Since news media play a central role in information dissemination across many culturally and linguistically diverse societies, enabling inclusive and equitable news recommendation requires addressing multilinguality as a core challenge (**C4**). Yet, despite billions of news articles being filtered and ranked by recommendation algorithms every day, personalization remains flawed or entirely absent for the majority of the world's languages.

Digital news platforms rely on recommendation systems to meet the diverse information needs of readers. However, most research focuses on major, resource-rich languages, overlooking the linguistic diversity of online communities. Moreover, existing work typically assumes monolingual news consumption, neglecting polyglot users, and resulting in a lack of multilingual benchmarks for developing recommenders suited to multilingual and low-resource contexts.

To address this gap, in this Chapter we introduce xMIND, an open, multilingual news recommendation dataset created by machine translating the English MIND dataset into 14 linguistically and geographically diverse languages with varying digital footprints. Using xMIND, we systematically evaluate several content-based neural news recommenders in zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer, examining both monolingual and bilingual consumption patterns. In FS-XLT we compare random and category-based replacement methods for incorporating target-language data during training. Our results show that (i) current NNRs, grounded in multilingual language models, experience significant performance drops in ZS-XLT, and (ii) injecting target-language data in FS-XLT provides limited improvements, especially for bilingual consumption. Notably, randomly injecting target-language news during training leads to greater performance gains

[*]This Chapter is adapted from: (1) **Andreea Iana**, Goran Glavaš, and Heiko Paulheim. 2024d. MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 553–563, New York, NY, USA. Association for Computing Machinery and (2) **Andreea Iana**, Goran Glavaš, and Heiko Paulheim. Multilinguality in MIND: Advancing Cross-Lingual News Recommendation with a Multilingual Dataset. 2025. *ACM Transactions on Recommender Systems*.

compared to category-based replacements. Our in-depth analysis of representation alignment between source and target languages within the language model shows that FS-XLT improves cross-lingual alignment primarily for high-resource languages, while low-resource languages remain weakly aligned with English. These findings highlight the need for broader research efforts in multilingual and cross-lingual news recommendation.

8.1 Introduction

The shift from traditional to digital news consumption has transformed news platforms into the primary medium of information for Internet users. This, in turn, propelled personalized news recommendation systems into the main tool used by news websites to address the individual information needs of readers worldwide. With the Internet's global reach, the linguistic diversity of its user base has significantly increased (Zuckerman, 2008; Ling et al., 2020). A substantial portion of these users are polyglots, consuming news in two or more languages. For instance, 22% of Americans speak a language other than English at home[1], whereas 65% of the working-age adults in the European Union are proficient in at least one foreign language.[2] This pattern is especially pronounced among migrant users, who report consuming media both in the host country's language and in their native languages, using them to facilitate integration while maintaining a connection to their home country (Naseer et al., 2024a; Alencar and Deuze, 2017; Bozdağ and Karakasoğlu, 2024; Elis, 2025). Despite this, most online content continues to be dominated by a handful of resource-rich languages, with English alone comprising over half of the world's digital text. Meanwhile, the vast majority of languages worldwide account for less than 0.1% of websites or are missing online representation altogether.[3]

News media serve as a cornerstone of democratic societies by keeping the public informed and providing a platform for sharing and debating ideas and opinions (Balkin, 2017; Helberger, 2021). In this context, recommender systems shape people's worldviews and opinions through the way in which they filter and propagate news (Pariser, 2011). On the one hand, research on personalized neural news recommenders (Wu et al., 2023) has predominantly centered on enhancing recommendation accuracy and diversity (Wu et al., 2020c, 2022c,h; Shi et al., 2022), by mitigating technical challenges in news encoding (Wang et al., 2018b; Wu et al., 2019b,e,a,c; Liu et al., 2020a; Qi et al., 2021a; Wu et al., 2021b; Yu et al., 2022; Jiang, 2023; Li et al., 2023d; Zhang and Wang, 2023) and user modeling (An et al., 2019; Qi et al., 2021c,e, 2022b; Wang et al., 2022c; Li et al., 2022; Qi et al., 2022a, Chapter 4). On the other hand, recent advances in neural machine translation (NMT) (Dabre et al., 2020; Fan et al., 2021; Haddow et al., 2022; Team et al., 2022; Kudugunta et al., 2023) and multilingual pretrained language models (mPLMs) (Conneau and Lample, 2019; Conneau et al., 2020; Wei et al., 2020; Xue et al., 2021; Workshop et al., 2022; Touvron et al., 2023; Wei et al., 2023) for low(er)-resource languages have started to democratize access to information for underrepresented communities (Joshi et al., 2020).

Nonetheless, two main limitations persist in the existing body of research on news recommendation. First, there is a scarcity of publicly available, diverse, multilingual news recommendation datasets that could be leveraged to develop efficient multilingual news recommenders and support effective cross-lingual transfer to resource-lean languages. Although the availability of adequate datasets is paramount for developing high-quality recommenders (e.g., see the Amazon dataset[4] for product recommendation, or MovieLens (Harper and Konstan, 2015) and Netflix (Bennett et al., 2007) for movie recommendation), the vast majority of news recommendation benchmarks remain monolingual (Kille et al., 2013; Gulla et al., 2017; de Koning et al., 2018; Gabriel De Souza et al., 2019; Lucas et al., 2023; Wu et al., 2020e). Furthermore, the few existent multilingual benchmarks

[1] https://data.census.gov/table/ACSST1Y2022.S1601?q=language
[2] https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180926-1
[3] https://w3techs.com/technologies/overview/content_language
[4] https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews

mostly feature high-resource languages, and are limited by narrow focus, small size (see Section 6.5), or proprietary constraints (Wu et al., 2021b). Second, the design of news recommenders for multi- and cross-lingual settings has been left largely unexplored. Traditional news recommenders typically cater to single-language usage, hindering the ability to browse and recommend content across multiple languages simultaneously. As a result, multilingual users often receive recommendations that are less relevant, unbalanced, and lacking in diversity (Ling et al., 2020). Overall, these limitations pose significant challenges for online news readers who are multilingual or consume news in resource-lean and/or underrepresented languages. Accounting for such multilingual consumption patterns is crucial for developing news recommender systems that serve minority groups fairly, as overlooking their linguistic practices risks reinforcing digital inequities.

Contributions. We address the above research gaps by introducing xMIND, a new large-scale and publicly available multilingual news dataset for multi- and cross-lingual recommendation. xMIND is constructed by translating the articles of the English-only dataset MIND (Wu et al., 2020e) into a diverse selection of 14 high- and low-resource languages, spanning five geographical macro-areas, and 13 distinct language families, leveraging the NLLB open-source machine translation system (Team et al., 2022). Compared to existing multilingual news recommendation datasets, xMIND is: **(1)** much more *diverse* – we include both resource-rich and resource-poor languages, covering a wide variety of geographical regions, language families, and scripts, with some of the languages being out-of-sample for existing mPLMs (i.e., not present in the mPLM's pretraining); **(2)** *parallel* – the same set of news has been translated into all target languages, enabling the direct comparison of the performance of multilingual news recommenders and cross-lingual transfer approaches across target languages; **(3)** *open source* – we release the dataset in the tab-separated format and provide scripts for loading and combining the news with the corresponding click logs from the MIND dataset in the NewsRecLib (see § 3.3.2) library.

We use xMIND to benchmark a range of state-of-the-art neural news recommenders, in ZS-XLT and FS-XLT setups, covering two realistic news consumption settings: monolingual and bilingual. Furthermore, we examine two methods for injecting target-language data during model training, and for creating bilingual news consumption patterns: (i) random replacement and (ii) category-based replacement of English news with their target-language translations from xMIND. We show that recommenders trained monolingually on English news suffer significant performance drops when evaluated on the target languages under ZS-XLT conditions, even when paired with a massively multilingual language model. More importantly, we demonstrate that target-language injection during training has a limited effect in mitigating these performance drops. This is particularly the case when adopting a category-based injection strategy in FS-XLT. Additionally, we show that fine-tuning the recommender's underlying language model with low-rank adapters offers a lightweight alternative that drastically reduces the number of trainable parameters while maintaining comparable performance. We further analyze cross-lingual alignment between English and target languages within the news encoder, finding that few-shot target language injection primarily benefits high-resource languages, whereas low-resource languages require substantially more target-language examples to achieve alignment that nonetheless remains weaker with English. These findings reveal the urgent need for developing more accurate and robust cross-lingual news recommendation approaches. Lastly, we assess the quality of the translations in xMIND through a human-based annotation task and comparison against translations obtained with a commercial NMT system.

8.2 Related Work

8.2.1 News Recommendation

A significant body of work has sought to improve personalization by enhancing the NNRs' core components, namely news and user modeling. Nonetheless, the vast majority of efforts have been

nearly exclusively deployed in monolingual settings. More specifically, despite the abundance of polyglot news readers, few works explore the behavior of NNRs in a multi- or cross-lingual scenario. Wu et al. (2021b) suggested instantiating the news encoder with mPLMs to enable news recommendation in diverse languages. Guo et al. (2023) proposed a novel news encoder based on an unsupervised cross-lingual transfer model to address the few-shot recommendation problem between record-rich and unpopular or early-stage recommendation platforms without overlapping users and with news in different languages. However, these works focus exclusively on (i) resource-rich languages and (ii) monolingual news consumption. News recommendation for multilingual news consumers, especially speakers of under-resourced languages, thus remains largely unexplored.

8.2.2 LLM-based Recommendation

More recently, the emergence of large language models (LLMs) and their remarkable natural language understanding and generation capabilities have significantly reshaped the design of recommender systems. Broadly, we distinguish two main paradigms. The first employs (L)LMs for a discriminative task, where the model functions primarily as a powerful encoder to capture semantic representations of users, items, or contexts (Lyu et al., 2024a; Wu et al., 2024b; Korikov et al., 2024). The second leverages LLMs generatively by producing recommendations directly in natural language, generating explanations for recommendations, or reformulating user queries to improve retrieval (Deldjoo et al., 2024a). Within the domain of news recommendation, in particular, a growing body of research is exploring how LLMs can be employed to enhance personalization, improve interpretability, and address challenges such as cold-start and dynamic user interests. Li et al. (2023b) evaluate ChatGPT for news recommendation, focusing on personalization, provider fairness, and fake news, while Li et al. (2023c) study fine-tuning strategies and show that although ChatGPT underperforms popularity-based recommenders in zero-shot settings, fine-tuning improves performance for users with consistent topical histories with the candidates but remains weak for those whose histories are not topically aligned with the candidate articles. Zhang and Wang (2023) propose Prompt4NR, framing the click prediction task as a cloze-style mask-prediction task, and finding that combining discrete, continuous, and hybrid prompts performs best. PGNR (Li et al., 2024) treats recommendation as a text-to-text language generation task with LLMs, while Yada and Yamana (2024) use LLMs to generate category descriptions of news articles to enhance news recommendation. Gao et al. (2024) introduce a generative news recommendation paradigm that leverages LLM knowledge to construct theme-level representations of news and users. EmbSum (Zhang et al., 2024b) generates user-interest summaries with LLMs to model long histories, whereas SPAR (Zhang et al., 2024a) uses LLMs to extract global interests from the user's history for user profile enrichment. ONCE (Liu et al., 2024b) examines both open- and closed-source LLMs to augment training data and improve content representation, and Zhao et al. (2025b) extensively compare shallow and large language models for news recommendation. Despite the increasingly multilingual capabilities of LLMs, all existing approaches are evaluated solely on monolingual datasets, e.g., MIND (Wu et al., 2020e), and overlook multilingual consumption patterns.

8.2.3 Recommendation Datasets

The advancement of recommender systems heavily depends on the existence and availability of suitable datasets. In the past decade, several public monolingual datasets have been proposed for training and benchmarking news recommenders: Plista (German, Kille et al., 2013), Adressa (Norwegian, Gulla et al., 2017), Globo (Portuguese, de Koning et al., 2018; Gabriel De Souza et al., 2019) and its recent improved version NPR (Lucas et al., 2023), MIND (English, Wu et al., 2020e), and the most recent EB-NeRD (Danish, Kruse et al., 2024). Among these, the MIND dataset has become a reference benchmark for the news recommendation community, given the limitations of the earlier datasets, such as a lack of original news texts, metadata information, or limited dataset

size (Wu et al., 2020e). However, these datasets consists only of monolingual news, and therefore, hinder the development of multilingual recommender systems. In Chapter 6, we have aimed to address this problem by proposing NeMig, a multilingual news recommendation dataset in English and German. NeMig contains articles on the topic of refugees and migration collected from German and US media outlets, and rich user data encompassing both click logs and demographic and political information. Besides covering only two major languages, NeMig is small (7K German and 10K English articles) and covers only one specific topic. Wu et al. (2021b) mention the multilingual news recommendation dataset collected from the MSN News platform to analyze the effectiveness of mPLM-based news encoders in multilingual news recommendations. Their dataset contains user data from seven countries (US, Germany, France, Italy, Japan, Spain, and Korea). Besides all seven included languages being very highly resourced, the dataset is proprietary (i.e., not publicly available).

8.2.4 Cold-Start Recommendation

Personalized news recommendation relies on user-item interactions for training, making it vulnerable to the cold-start recommendation problem when data on new items or users is sparse or unavailable (Gope and Jain, 2017). To address this, prior work has explored meta-learning (Lee et al., 2019; Zhu et al., 2021b; Luo et al., 2025), knowledge distillation (Wang et al., 2021b; Huang et al., 2023), or contrastive learning (Wei et al., 2021; Zhou et al., 2023), to transfer information from warm to cold items, thereby improving predictions on unseen cases. Other approaches generate synthetic interactions for cold-start items from behavioral signals (Chen et al., 2022; Liu et al., 2024c; Huang et al., 2025a). More recently, LLMs have been employed to tackle cold-start challenges, leveraging their strong zero-shot and few-shot capabilities (Sanner et al., 2023; Wang and Lim, 2023; Hou et al., 2024; Wu et al., 2024c; Zhang et al., 2025b). Retrieval-augmented generation methods have also emerged, enhancing LLM-based recommenders with external knowledge sources (Di Palma, 2023; Wu et al., 2024a; Contal and McGoldrick, 2024; Kieu et al., 2025b). The news recommendation domain has followed similar strategies. For example, Alshehri and Zhang (2022) propose a generative adversarial zero-shot learning framework that constructs virtual behavioral representations for cold users or items, whereas Li et al. (2023a) introduce a meta-learning approach that factorizes user preferences into time-specific and time-shift representations. GZRec (Hassan et al., 2024) synthesizes cold- and warm-start interactions for users and news, and frames cold-start recommendations as zero-shot predictions. BCE4ZSR (Rauf et al., 2024b) and ZS-CEBE (Rauf et al., 2024a) both adopt a knowledge distillation framework, where a cross-encoder models user-news interactions and a bi-encoder generates user and item embeddings. However, these methods are generally evaluated only on monolingual datasets, e.g., MIND (Wu et al., 2020e), Adressa (Gulla et al., 2017), and overlook language transfer or language-dependent user preferences. To the best of our knowledge, Guo et al. (2023) are the only ones to study cold-start in news recommendation as a few-shot recommendation problem, showing that cross-lingual transfer, namely transferring user-news preferences from a many-shot source to a few-shot target domain, can improve performance. Cross-lingual news recommendation constitutes a specific case of cold-start: (i) new articles may appear in a language unseen by the recommender, and (ii) users histories may span different or multiple languages. In other words, users or items in one language are effectively cold-start cases relative to a knowledge-rich language familiar to the recommender. Yet, the field still lacks multilingual benchmark datasets to evaluate and advance methods that tackle the cold-start challenges arising from linguistic diversity.

8.3 xMIND Dataset

We create xMIND with two primary considerations in mind: (1) covering languages that are mutually *diverse* linguistically, geographically, and in terms of amount of available text corpora (i.e., high or low resource) and (2) creating a multilingual news dataset that is multi-parallel, i.e., where an

article (i.e., or its translation) exists in each covered language. The former allows for a more realistic estimate of global multilingual and cross-lingual performance of news recommendation models (Joshi et al., 2020; Ponti et al., 2020), whereas the latter enables direct comparability of recommenders' performance across target languages. We thus create xMIND by translating 130,379 unique news articles from the train, development, and test portions of the English MIND dataset (Wu et al., 2020e) (i.e., union of MINDlarge and MINDsmall) into 14 different languages using the NLLB 3.3B open-source NMT model (Team et al., 2022). The MIND news articles consist of a title and an abstract, and are additionally annotated with the topical category and Wikipedia-disambiguated named entities extracted from the title and abstract.[5] Note that, while we translate only the title and abstract of each news, these can still be combined with the corresponding linked named entities for usage in knowledge-aware recommendation models (Iana et al., 2024).

Language Selection. We select target languages for xMIND based on the following criteria: (1) linguistic diversity in terms of typological properties (Dryer and Haspelmath, 2013; Littell et al., 2017), language family, and geographical provenance, (2) script diversity, (3) amount of available language resources, primarily raw corpora (i.e., inclusion of both high- and low-resource languages), and (4) coverage by NLLB (Team et al., 2022). Table 8.1 lists the selected languages, summarizing the following information, in accordance with the #BenderRule (Bender, 2019):

- **Code**: The three-letter ISO 693-3 code of the language;
- **Language**: In case of multiple denominations, we use the language name from the World Atlas of Structures (WALS) (Dryer and Haspelmath, 2013). We cross-reference the names with two other major linguistic resources, Glottolog (Hammarström et al., 2021) and Ethnologue (Lewis et al., 2009);
- **Script**: We provide the English name of the script;
- **Family and Genus**: Language family and genus from WALS (Dryer and Haspelmath, 2013) and Glottolog (Hammarström et al., 2021);
- **Resource Level**: We borrow NLLB's (Team et al., 2022) classification of languages into *low-* and *high-resource*;
- **Total Speakers**: We report the total number of speakers of the language, including L1-level (first-language) and L2-level (second-language) speakers, according to Ethnologue.[6]

We follow Ponti et al. (2020) and compute three different diversity scores for our language sample: (i) typology index, (ii) family index, and (iii) geographical index. **1)** The *typology index* is based on 103 typological binary features of each language from URIEL (Littell et al., 2017): each feature indicates the presence or absence of a particular linguistic property in a language. As per Ponti et al. (2020), we compute the typology index as the average of entropy scores computed independently for each feature;[7] **2)** The *family index* is the number of distinct language families divided by the sample size; **3)** The *geography index* is the entropy of the distribution of languages in the sample over 6 geographic macro-areas of the world.[8]

Table 8.2 reports the three metrics for xMIND, as well as for NeMig (cf. Section 6.5) and the proprietary dataset from Wu et al. (2021b) (dubbed *Wu et al.*). xMIND offers the most diverse sample in terms of all diversity indices. The sample of languages spans five out of the six macro-areas, and

[5]Note that 5.4% of the news in the entire MIND dataset do not contain an abstract.

[6]We use the latest statistics available in January 2024 at https://www.ethnologue.com/.

[7]The entropy of a feature for which all languages in the sample have the same value is 0; the entropy has the maximal value ($\log 2$) if the feature is present for the same number of languages as for which it is absent.

[8]The six macro-areas, as defined by Dryer and Haspelmath (2013), are: Africa, Australia, Eurasia, North America, Papunesia, and South America.

Code	Language	Script	Macro-area	Family	Genus	Total Speakers (M)	Res.
SWH	Swahili	Latin	Africa	Niger-Congo	Bantu	71.6	high
SOM	Somali	Latin	Africa	Afro-Asiatic	Lowland East Cushitic	22.0	low
CMN	Mandarin Chinese	Han	Eurasia	Sino-Tibetan	Sinitic	1,138.2	high
JPN	Japanese	Japanese	Eurasia	Japonic	Japanesic	123.5	high
TUR	Turkish	Latin	Eurasia	Altaic	Turkic	90.0	high
TAM	Tamil	Tamil	Eurasia	Dravidian	Dravidian	86.6	low
VIE	Vietnamese	Latin	Eurasia	Austro-Asiatic	Vietic	85.8	high
THA	Thai	Thai	Eurasia	Tai-Kadai	Kam-Tai	60.8	high
RON	Romanian	Latin	Eurasia	Indo-European	Romance	24.5	high
FIN	Finnish	Latin	Eurasia	Uralic	Finnic	5.6	high
KAT	Georgian	Georgian	Eurasia	Kartvelian	Georgian-Zan	3.9	low
HAT	Haitian Creole	Latin	North America	Indo-European	Creoles and Pidgins	13.0	low
IND	Indonesian	Latin	Papunesia	Austronesian	Malayo-Sumbawan	199.1	high
GRN	Guarani	Latin	South America	Tupian	Maweti-Guarani	(L1 only) 6.7	low

Table 8.1: The 14 languages of xMIND. We display the language *Code* (ISO 693-3), language name, *Script*, *Macro-area*, and language *Family* and *Genus*. *Res.* indicates whether the language is classified as high or low-resource according to Team et al. (2022).

	Range	xMIND	NeMig	Wu et al.
Typology	[0, 1]	0.42	0.05	0.31
Family	[0, 1]	0.93	0.50	0.43
Geography	[0, ln 6]	1.13	0.00	0.00

Table 8.2: Indices of typological, genealogical, and geographical diversity for the language samples of different multilingual news recommendation datasets.

13 distinct language families covering 14 different genera. We excluded languages from Australia, as they (i) have an extremely low number of native speakers (i.e., at most spoken by a few thousand people), and (ii) are not supported by NLLB (Team et al., 2022). Additionally, GRN and HAT (i) are spoken in South and North America, both originating from underrepresented macro-areas, and (ii) have not been seen in pretraining of XLM-RoBERTa (Conneau et al., 2020). Moreover, xMIND covers five low-resource languages (Somali, Tamil, Georgian, Haitian Creole, and Guarani) and six different scripts: Latin and Georgian are *alphabet* scripts; Japanese and Chinese Han are *logographic* scripts, whereas the Tamil and Thai are written in *Abugida* script type.

NLLB: Hyperparameter Tuning. We tune the hyperparameters of the NLLB translation model (Team et al., 2022) using a subset of Global Voices (GV, Tiedemann, 2012) as validation set.[9] GV constitutes a parallel corpus of news stories in 46 languages collected from the Global Voices website.[10] We construct the validation dataset by selecting the data files for all covered pairs of *English* (ENG) as the source and any of the xMIND languages in xMIND as the target: this results in six language pairs, statistics of which are reported in Table 8.3. We compare four decoding strategies: greedy, multinomial sampling, beam-search, and beam search with multinomial sampling. For the beam-search decoding strategies, we search for the optimal number of beams in the range $[2, 8]$ and use default values for all other hyperparameters. We evaluate the translation quality using the sacreBLEU score (Post, 2018). With the goal of finding the best decoding strategy for a broad range of languages, we compute the macro-average over the six language pairs in our validation dataset. As shown in Table 8.4, we identify beam search decoding with 4 beams as the best choice: using more beams increases computational cost while bringing negligible sacreBLEU gains.

[9]We use the most recent version available online in October 2023, namely *GlobalVoices v2018q4*. The original data can be accessed at https://opus.nlpl.eu/GlobalVoices/corpus/version/GlobalVoices.

[10]https://globalvoices.org/

Language Pair	Sentence pairs	Words (M)
ENG -> CMN	137,737	2.83
ENG -> SWH	30,338	1.13
ENG -> IND	15,266	0.54
ENG -> JPN	8,595	0.18
ENG -> TUR	7,479	0.24
ENG -> RON	4,265	0.17

Table 8.3: Statistics of the subset of Global Voices (Tiedemann, 2012) used as validation data for tuning the NLLB (Team et al., 2022) hyperparameters. We report the number of sentence pairs and of words (in millions).

Decoding Strategy	# Beams	sacreBLEU
Greedy	1	18.42
Multinomial sampling	1	11.97
Beam Search	4	19.03
Beam Search Multinomial Sampling	4	18.87

Table 8.4: Hyperparameter optimization results on the subset of Global Voices (Tiedemann, 2012). We report only the results obtained with the best number of beams. We report macro-average sacreBLEU scores over six language pairs.

Final Dataset. The xMIND dataset contains 130,379 unique news in the 14 different languages listed in Table 8.1. Each article contains a news ID, a translated title, and a translated abstract – if one was provided in the corresponding English article from MIND (Wu et al., 2020e). Following Wu et al. (2020e), we split xMIND, for each language, firstly into a small and a large version of the dataset, and secondly, into train, development, and test portions, each corresponding to the original splits of the MIND dataset. Table 8.5 lists the number of articles in each variant and split of the dataset. We release xMIND publicly, in tab-separated format at https://github.com/andreeaiana/xMIND. xMIND can be combined with additional news and behavioral information provided in MIND (Wu et al., 2020e), using the news IDs. Additionally, to facilitate a seamless integration with existing NNRs, we implement the data loading functionality for xMIND in NewsRecLib (see § 3.3.2). Since xMIND can be used beyond cross-lingual news recommendation in text retrieval or translation tasks, we also release our dataset in Parquet format, in both the small and the large version, on HuggingFace Datasets.[11]

Small		Large		
Train	Dev	Train	Dev	Test
51,282	42,416	101,527	72,023	120,959

Table 8.5: Number of news in the different splits of xMIND.

8.4 Experimental Setup

We systematically benchmark a range of state-of-the-art NNRs in ZS-XLT and FS-XLT settings. Our experiments encompass two types of news consumption patterns: monolingual and bilingual.

[11]The small version of xMIND can be downloaded from https://huggingface.co/datasets/aiana94/xMINDsmall. The large version of the dataset can be downloaded from https://huggingface.co/datasets/aiana94/xMINDlarge.

8.4.1 Benchmarked Neural News Recommenders

We evaluate several content-based NNRs: (1) *NAML* (Wu et al., 2019a), (2) *LSTUR* (An et al., 2019), (3) *MINS* (Wang et al., 2022c), (4) *CAUM* (Qi et al., 2022b), (5) *TANR* (Wu et al., 2019c), (5) *MINER* (Li et al., 2022), and (6) *MANNeR* (cf. Chapter 7) (only the `CR-Module` responsible for pure content-based recommendation, without any aspect-based personalization or diversification). Additionally, we include (7) NAML_{CAT} as a baseline—a language-agnostic variant of NAML that generates news embeddings solely from randomly initialized category vectors, and derives user representations by attending to the embeddings of clicked news articles. With the exception of MINER and MANNeR, which are designed with a PLM-based news encoder, the remaining NNRs originally contextualize word embeddings using CNNs (Kim, 2014), additive attention (Bahdanau et al., 2015), or multi-head self-attention (Vaswani et al., 2017) networks. For fair comparison and to enable multilingual recommendations, we follow Wu et al. (2021b), and replace the original news encoders of these NNRs with an mPLM. NAML, LSTUR, MINS, TANR, and CAUM leverage category information in addition to the news text, whereas CAUM and MANNeR also encode named entities. Models with multiple input features either concatenate (i.e, LSTUR, CAUM) or attend over them (i.e, NAML, MINS, MANNeR) to produce the final news embedding.

The recommenders further vary in their user encoders: NAML and TANR encode users' preferences using additive attention; MINS combines multi-head self-attention with a multi-channel GRU-based (Cho et al., 2014) recurrent network and additive attention; MINER introduces a poly-attention approach based on multiple additive attentions to learn various interest vectors for each user. Additionally, LSTUR and CAUM differentiate between short and long-term user preferences. More specifically, LSTUR encodes the former from the clicked news embeddings with a GRU, and the latter via randomly initialized and fine-tuned embeddings; the final user representation is produced by combining the two embeddings.[12] CAUM models long-term dependencies between clicked news with a candidate-aware self-attention network, short-term user interests from adjacent clicks with a candidate-aware CNN, and obtains the final candidate-aware user embedding by attending over the two intermediate representations. In contrast to the other models, MANNeR does not use a parameterized user encoder. Instead, it uses a late fusion approach that involves mean pooling of dot-product scores between each of the candidates and the clicked news.

All benchmarked models compute recommendation scores as the dot product between the candidate representations and the clicked news. MANNeR is optimized using a supervised contrastive loss (Khosla et al., 2020), whereas the other models are trained by minimizing the standard cross-entropy loss.

8.4.2 Training and Optimization Details

Data. We combine the xMIND news with the corresponding click logs and additional news annotations (i.e., categories and named entities) from MIND based on the news IDs. We conduct all experiments on the *small variant* of the resulting dataset (i.e., MINDsmall). For details on the corresponding dataset statistics, we refer the reader to § 3.2.2.

Training Details. We conduct all the experiments with XLM-RoBERTa (Conneau et al., 2020) as the underlying language model in the news encoder of all recommenders. In Section 8.5, we experiment with two additional multilingual language models, namely LaBSE (Feng et al., 2022) and umT5 (Chung et al., 2023). Appendix F.1 indicates which languages in xMIND are already included in the pretraining corpora of these mPLMs. We fine-tune only the last four layers of all mPLMs.[13] We use 100-dimensional TransE embeddings (Bordes et al., 2013), pretrained on Wikidata, to initialize the

[12] Note that in our experiments, we use the *ini* strategy of LSTUR for obtaining the final user embedding, as it outperforms the *con* variant in preliminary evaluations. We refer the reader to An et al. (2019) for more details.

[13] In the interest of computational efficiency, we keep the bottom eight layers of the Transformer encoders frozen.

entity encoder in the news encoder of the knowledge-aware models.[14] In line with prior work, we set the maximum history length to 50 and sample four negatives per positive sample during training, as per Wu et al. (2022b). We tune the main hyperparameters of all NNRs. We train all recommenders with a news encoder based on XLM-RoBERTa or LaBSE with mixed precision and a batch size of 8, while those using umT5 are trained with float 32 precision and a batch size of 4. We train all NNRs with a learning rate of 1e-5, for 10 epochs with early stopping and optimizing with the Adam algorithm (Kingma and Ba, 2014). We refer the reader to Appendix F.2 for further details concerning the hyperparameter settings. We repeat each experiment three times, and report averages and standard deviations for the standard metrics: AUC, MRR, NDCG@10.[15] To ensure comparability, we train and evaluate all NNRs using the NewsRecLib (cf. § 3.3.2) library.[16]

Infrastructure and Compute. We train all models on a cluster with virtual machines, on single NVIDIA A100 40/80 GB GPUs or A40 48 GB GPUs.

8.4.3 Cross-Lingual Recommendation Scenarios

We benchmark the NNRs in two evaluation setups: (i) **zero-shot (ZS-XLT)** and (ii) **few-shot (FS-XLT)** cross-lingual recommendation. Firstly, through ZS-XLT we aim to investigate the capabilities of NNRs trained monolingually in English (i.e., on the MIND news) to generate recommendations in another language (i.e., in one of the 14 languages in xMIND). Under ZS-XLT, the user history and candidates during training are *monolingual*, in English only. Secondly, with FS-XLT we seek to determine whether target language injection during training benefits the models' performance compared to pure ZS-XLT. In the FS-XLT setting, we explore two approaches for substituting English news with target-language news: *random* and *category-based*. In the *random* variant, we progressively replace between 10% and 90% of the English news (in both user histories and candidate sets) with target-language articles. In the *category-based* variant, we replace the English news in each user's top-k most-clicked categories with their target-language counterparts, again in both user histories and candidate sets. We test $k \in \{1, 2, 3\}$, corresponding to the one, two, or three most-clicked categories per user. For a fair setup (i.e., no knowledge of the test data distributions during training), the distribution of languages in our validation sets mirror the language ratios of the respective training sets (Schmidt et al., 2022).

We couple the two training settings (monolingual and bilingual), with two corresponding types of *news consumption patterns* during inference: (i) **monolingual** (denoted `MONO`) – the user reads news and receives suggestions only in the target language, and (ii) **bilingual** (denoted `BILING`) – the user consumes news in English and in another language, and recommendations are also provided in the same two languages. To construct the bilingual user history, and candidate set, respectively, we again consider a *random* and a *category-based* approach. Under the random approach, we randomly replace a subset of the English news items with their corresponding xMIND translations in the target language. Following the procedure used in bilingual training, we vary the portion of replaced news from 10% to 90% in increments of 10%. In the category-based replacement setting, we assume (i) users typically prefer consuming their favorite topics in their primary language, and (ii) their interests vary, leading to different top-clicked categories. Hence, we replace the English news associated with each user's top k most frequently clicked categories with the corresponding target-language translations from xMIND. Similar to our FS-XLT setup, we incrementally replace up to three top-clicked categories per user.

The two training setups, each combined with both consumption patterns, thus result in four types of experiments: **(i) ZS-XLT$_{\texttt{MONO}}$** – monolingual training (in English) and evaluation on monolingual news consumption in the target language; **(ii) ZS-XLT$_{\texttt{BILING}}$** – monolingual training (in English) and

[14] The entity vectors are provided as part of the original MIND dataset (Wu et al., 2020e).
[15] For brevity, we omit results for NDCG@5, as they exhibit the same patters as NDCG@10.
[16] https://github.com/andreeaiana/newsreclib

mPLM	Model	# Parameters (M)		AUC			MRR			NDCG@10		
		Trainable	Total	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ
–	NAML_{CAT}	0.387	0.387	$55.46_{\pm 0.18}$		0.0	$31.12_{\pm 0.56}$		0.0	$35.81_{\pm 0.59}$		0.0
XLM-RoBERTa	CAUM	227	284	$57.82_{\pm 3.01}$	$55.90_{\pm 1.75}$	-3.32	$32.92_{\pm 1.68}$	$31.38_{\pm 1.62}$	-4.68	$37.49_{\pm 1.71}$	$35.96_{\pm 1.58}$	-4.08
	LSTUR	313	370	$56.80_{\pm 1.36}$	$56.28_{\pm 1.68}$	-0.92	$33.00_{\pm 0.59}$	$31.53_{\pm 0.85}$	-4.47	$37.45_{\pm 0.54}$	$36.03_{\pm 0.85}$	-3.78
	MANNeR	222	279	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	0.00	$35.58_{\pm 0.31}$	$33.03_{\pm 0.54}$	-7.15	$40.17_{\pm 0.21}$	$37.64_{\pm 0.44}$	-6.28
	MINER	221	278	$57.73_{\pm 7.33}$	$55.81_{\pm 4.33}$	-3.32	$31.71_{\pm 4.95}$	$30.20_{\pm 3.42}$	-4.76	$36.45_{\pm 4.84}$	$35.02_{\pm 3.51}$	-3.90
	MINS	226	283	$59.89_{\pm 0.29}$	$56.94_{\pm 1.40}$	-4.93	$34.7_{5\pm 0.24}$	$33.11_{\pm 0.51}$	-4.70	$39.35_{\pm 0.20}$	$37.64_{\pm 0.50}$	-4.35
	NAML	224	280	$52.85_{\pm 2.27}$	$52.49_{\pm 2.60}$	-0.68	$35.98_{\pm 0.44}$	$33.98_{\pm 0.95}$	-5.56	$40.43_{\pm 0.39}$	$38.38_{\pm 1.02}$	-5.06
	TANR	223	280	$54.18_{\pm 5.91}$	$53.27_{\pm 1.91}$	-1.68	$35.47_{\pm 0.95}$	$32.14_{\pm 0.90}$	-9.40	$40.03_{\pm 0.86}$	$36.78_{\pm 0.88}$	-8.11
LaBSE	CAUM	414	471	$64.92_{\pm 0.83}$	$63.52_{\pm 0.80}$	-2.16	$34.36_{\pm 0.40}$	$33.53_{\pm 0.69}$	-2.69	$39.40_{\pm 0.21}$	$38.34_{\pm 0.54}$	-2.70
	LSTUR	504	560	$65.62_{\pm 0.74}$	$63.28_{\pm 0.88}$	-3.56	$33.82_{\pm 0.67}$	$32.73_{\pm 0.78}$	-3.22	$\mathbf{47.63}_{\pm \mathbf{0.50}}$	$\mathbf{46.71}_{\pm \mathbf{0.60}}$	-1.95
	MANNeR	415	472	$\mathbf{68.31}_{\pm \mathbf{1.26}}$	$\mathbf{66.89}_{\pm \mathbf{1.21}}$	-2.09	$36.52_{\pm 0.65}$	$35.24_{\pm 0.52}$	-3.50	$41.11_{\pm 0.65}$	$39.81_{\pm 0.48}$	-3.16
	MINER	414	471	$60.19_{\pm 4.61}$	$62.16_{\pm 0.44}$	+3.26	$35.82_{\pm 0.82}$	$33.37_{\pm 0.54}$	-6.84	$40.27_{\pm 0.92}$	$37.78_{\pm 0.44}$	-6.18
	MINS	416	473	$63.70_{\pm 1.02}$	$62.31_{\pm 0.77}$	-2.19	$34.31_{\pm 0.80}$	$33.58_{\pm 0.58}$	-2.11	$39.06_{\pm 0.69}$	$38.22_{\pm 0.46}$	-2.17
	NAML	414	471	$51.59_{\pm 0.71}$	$51.93_{\pm 0.65}$	+0.66	$36.63_{\pm 0.68}$	$\mathbf{35.91}_{\pm \mathbf{0.51}}$	-1.98	$41.25_{\pm 0.71}$	$40.42_{\pm 0.49}$	-1.99
	TANR	414	471	$63.93_{\pm 0.10}$	$62.40_{\pm 0.26}$	-2.39	$\mathbf{36.80}_{\pm \mathbf{0.18}}$	$34.99_{\pm 0.38}$	-4.91	$41.42_{\pm 0.18}$	$39.56_{\pm 0.39}$	-4.49
umT5	CAUM	203	288	$60.77_{\pm 0.93}$	$59.60_{\pm 0.83}$	-1.92	$33.41_{\pm 0.61}$	$32.18_{\pm 0.58}$	-3.69	$37.86_{\pm 0.47}$	$36.72_{\pm 0.48}$	-2.99
	LSTUR	289	374	$55.10_{\pm 1.32}$	$52.55_{\pm 1.62}$	-4.63	$31.69_{\pm 1.45}$	$30.18_{\pm 1.80}$	-4.77	$45.94_{\pm 1.06}$	$44.71_{\pm 1.33}$	-2.68
	MANNeR	198	283	$66.31_{\pm 0.49}$	$60.59_{\pm 1.18}$	-8.63	$35.13_{\pm 0.55}$	$31.48_{\pm 0.98}$	-10.39	$39.55_{\pm 0.38}$	$35.65_{\pm 0.88}$	-9.86
	MINER	197	282	$53.09_{\pm 0.71}$	$50.73_{\pm 0.30}$	-4.44	$26.85_{\pm 0.39}$	$25.11_{\pm 0.26}$	-6.49	$31.17_{\pm 0.43}$	$29.21_{\pm 0.22}$	-6.30
	MINS	202	287	$54.36_{\pm 1.82}$	$52.91_{\pm 1.25}$	-2.67	$30.10_{\pm 0.70}$	$28.95_{\pm 0.77}$	-3.82	$44.69_{\pm 0.51}$	$43.82_{\pm 0.57}$	-1.94
	NAML	199	284	$54.66_{\pm 0.53}$	$52.68_{\pm 0.62}$	-3.63	$31.04_{\pm 0.84}$	$30.01_{\pm 1.39}$	-3.32	$45.49_{\pm 0.66}$	$44.66_{\pm 1.06}$	-1.82
	TANR	199	284	$53.23_{\pm 0.48}$	$49.71_{\pm 0.47}$	-6.62	$28.33_{\pm 0.17}$	$25.96_{\pm 0.48}$	-10.39	$33.18_{\pm 0.14}$	$30.34_{\pm 0.45}$	-8.56

Table 8.6: $\text{ZS-XLT}_{\text{MONO}}$ recommendation performance. For each model, we report the number of model parameters (in millions), and the performance (i) on the English MIND dataset (denoted ENG), (ii) averaged across all 14 target languages in xMIND (denoted AVG), and (iii) the relative percentage difference between average $\text{ZS-XLT}_{\text{MONO}}$ and ENG performance (%Δ). We report averages and standard deviations across three runs. The best results per column are highlighted in bold, the second best are underlined.

evaluation on bilingual news consumption in English and the target language; **(iii)** $\textbf{FS-XLT}_{\textbf{MONO}}$ – bilingual training in a mixture of English and target language and evaluation on monolingual news consumption in the target language; **(iv)** $\textbf{FS-XLT}_{\textbf{BILING}}$ – bilingual training in a mixture of English and target-language and evaluation on bilingual news consumption in English and the target language.[17]

8.5 Zero-Shot Cross-lingual Transfer

We evaluate the performance of the benchmarked models under ZS-XLT, comparing their results to those obtained on the English data (MIND). We first discuss the performance of the models within the standard monolingual news consumption pattern, followed by an evaluation in the bilingual consumption setting. An ideal NNR should consistently rank relevant (positive) candidates higher than irrelevant (negative) ones, regardless of language. Consequently, our analysis primarily focuses on ranking performance, utilizing the NDCG@10 metric.

Monolingual Consumption Pattern. In most of our evaluations, we rely on XLM-RoBERTa Base (Conneau et al., 2020) – a multilingual, encoder-only Transformer model – as the main text encoder for all NNRs (see § 8.4.2 for details). We also compare the models' performance when coupled with two additional language models: (i) LaBSE (Feng et al., 2022), a BERT-based multilingual sentence encoder, and (ii) the encoder of umT5 (Chung et al., 2023), which follows the encoder-decoder architecture of the mT5 family (Xue et al., 2021) and is pretrained on multilingual corpora. Table 8.6 summarizes the $\text{ZS-XLT}_{\text{MONO}}$ recommendation performance for each model, averaged across the 14 languages in xMIND. These results are compared against the models' performance on MIND (i.e.,

[17] Note that due to the high computational costs, in the case of $\text{FS-XLT}_{\text{BILING}}$ with a random-replacement strategy, we replace the same percentage of English news with articles in the target language both during training and testing (e.g., 10% RON with 90% ENG during training results in the same mixture in testing). We adopt the same setup in the case of the category-based replacement in $\text{FS-XLT}_{\text{BILING}}$.



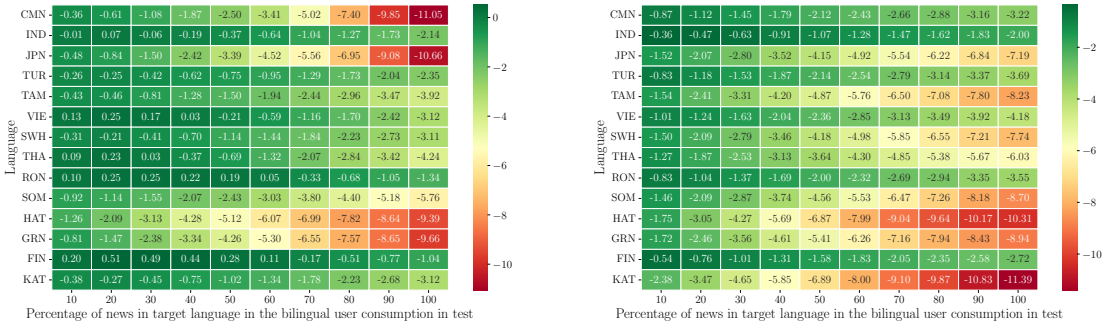
Figure 8.1: ZS-XLT$_{\text{MONO}}$ ranking performance (NDCG@10), across the 14 languages in xMIND and English, for three different types of mPLMs used in the recommenders' news encoder.

trained and tested solely in English). For additional context, we include NAML$_{\text{CAT}}$, which generates representations based only on topical categories rather than content. Because it is content- and language-agnostic, NAML$_{\text{CAT}}$ serves as a suitable baseline for both English and the target languages.

In the monolingual English setting on the MIND dataset, all XLM-RoBERTa- and LaBSE-based NNRs outperform the category-based NAML$_{\text{CAT}}$ baseline. Using XLM-RoBERTa, relative improvements in NDCG@10 range from 1.78% (MINER) to 12.91% (NAML), while LaBSE-based recommenders yield gains between 9.08% (MINS) and 33.01% (LSTUR). Generally, NNRs equipped with umT5 perform on par with or slightly better than those using XLM-RoBERTa, except for MINER and TANR, which do not surpass the content-agnostic baseline.

Under ZS-XLT$_{\text{MONO}}$, however, XLM-RoBERTa-based NNRs exhibit weaker performance relative to NAML$_{\text{CAT}}$, achieving at most a 7.19% improvement (NAML) or even a 2.19% drop (MINER). As in the English-only case, recommenders using LaBSE consistently surpass the content-agnostic NAML$_{\text{CAT}}$, with relative gains of up to 12.87% (NAML). Meanwhile, the ranking performance of umT5-based models varies substantially, exhibiting drops of up to 18.43% (MINER) or gains of up to 24.85% (LSTUR) compared to their fully English-trained and evaluated counterparts.

We also observe that XLM-RoBERTa-based NNRs trained in English and evaluated on the target languages in xMIND suffer average performance declines of 3.78% (LSTUR) to 8.11% (TANR) relative to their English results. While top-performing English models (e.g., NAML, MANNeR, TANR) still rank highly under ZS-XLT$_{\text{MONO}}$, they also experience the steepest relative drops. It is important to remember that a random recommender would exhibit the same performance across all languages (i.e., a 0% difference from English). As such, absolute performance in cross-lingual evaluation remains more crucial than the magnitude of these relative drops. Although using a sentence encoder does not entirely eliminate the gap between ZS-XLT and in-language performance, it does narrow it, reducing performance declines to between 1.95% (NAML) and 6.18% (MINER) compared to English.



(a) Results for NAML

(b) Results for MANNeR

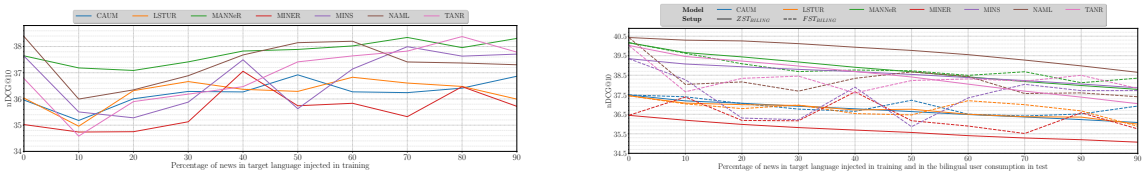
Figure 8.2: Relative percentage difference in ranking performance (NDCG@10), under ZS-XLT$_{\text{BILING}}$ compared to full English training and testing.

Fig. 8.1 provides a language-by-language breakdown of ZS-XLT$_{\text{MONO}}$ performance across the 14 target languages in xMIND. XLM-RoBERTa-based NNRs fare best on VIE, IND, RON, and FIN, which are well-represented in mPLM's pretraining corpus (Conneau et al., 2020). Conversely, performance is poorest across models for KAT, HAT, and GRN, the latter two being completely unseen during pretraining (though XLM-RoBERTa was at least exposed to the Latin script). All LaBSE-based recommenders exhibit notable drops on KAT, with some also performing worse on GRN (i.e., out-of-sample for LaBSE), THA, and CMN. Finally, results for umT5-based NNRs are less uniform: certain models underperform on CMN and SOM, but perform comparatively well on low-resource languages such as GRN and KAT.

Bilingual Consumption Pattern. We next examine how each model's ranking performance changes under bilingual user consumption (ZS-XLT$_{\text{BILING}}$), using XLM-RoBERTa as the base mPLM for all recommenders. In these experiments, we apply a *random replacement* of English news with their translations in the target language, both in user histories and candidate sets. Fig. 8.2 presents results for NAML and MANNeR – representative of the overall trends – relative to their respective English-only baselines, comparing (i) performance across different target languages and (ii) varying proportions of target-language news.[18] Overall, performance steadily declines as the proportion of target-language news increases. This drop is most pronounced for languages unseen in the mPLM's pretraining (i.e., HAT and GRN), as well as for KAT. Interestingly, even though all models share the same mPLM, they are not equally robust to the choice of target language. For instance, NAML's performance on CMN drops by up to 9.85% (with 90% of the user history in CMN), exceeding its losses on HAT and GRN, whereas MANNeR exhibits only a 3.16% decline under the same conditions. However, MANNeR suffers a larger decrease for HAT, a pattern that is more pronounced in TANR and LSTUR. Additionally, for certain languages and models (e.g., LSTUR with RON or FIN), the decline in ranking performance is smaller when either English or the target language dominates the user history.

Given the models' diverse architectural designs, this highlights the need to investigate the robustness of the architecture, particularly of the user encoder, under varying news consumption patterns, moving beyond the traditionally assumed monolingual user history. Our prior studies (cf. Chapters 4 and 5) have shown that NNRs based on sentence encoders or employing late fusion, rather than a parameterized user encoder, tend to exhibit better robustness. Nonetheless, these findings were derived exclusively from monolingual news consumption scenarios, and their validity may differ substantially in the context of multilingual user histories.

[18] We focus on NAML and MANNeR for clarity, as their behavior reflects that of other recommenders.



(a) FS-XLT$_{\text{MONO}}$ results

(b) FS-XLT$_{\text{BILING}}$ versus ZS-XLT$_{\text{BILING}}$ results

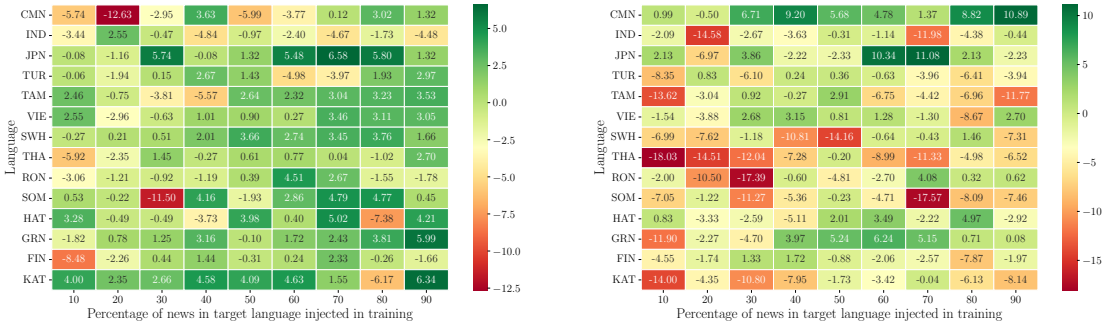
Figure 8.3: FS-XLT ranking performance (NDCG@10) averaged over the 14 languages in xMIND, for various target-language injection ratios during training, and in the user's bilingual news consumption in test. 0% denotes fully monolingual English training (i.e., ZS-XLT). We show the results for the *random replacement* strategy in FS-XLT.

8.6 Few-Shot Cross-lingual Transfer

Few-shot transfer, which requires the injection of a few target-language instances during model training, has been shown to yield sizable performance gains in NLP tasks (Schmidt et al., 2022). Therefore, it is often leveraged as an effective remedy to the dramatic performance drops suffered by multilingual models in ZS-XLT setups, particularly for resource-lean target languages that are linguistically distant from the source language (Lauscher et al., 2020). Motivated by these findings, we further analyze whether adding target-language data in training also benefits news recommenders. For all experiments in this section, we use only XLM-RoBERTa as the underlying language model in the news encoder of all NNRs. We first examine the effects of few-shot target language injection based on the random replacement strategy. Afterwards, we investigate how the category-based replacement in FS-XLT affects the NNR's training. Finally, we compare the two replacement methods against ZS-XLT in a monolingual news consumption setting.

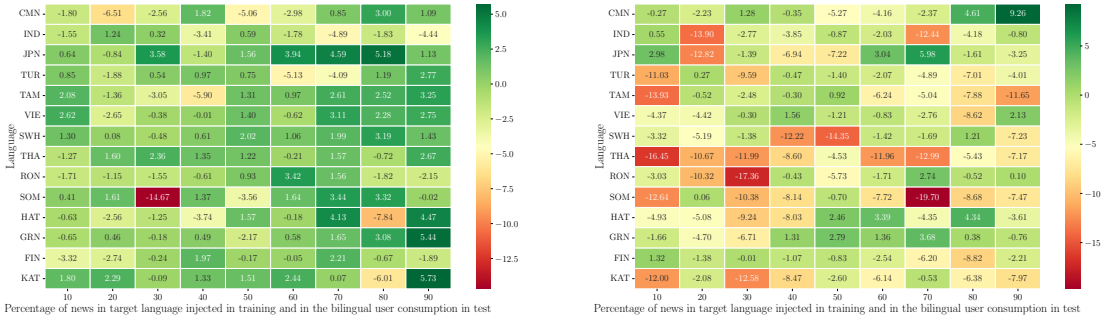
8.6.1 Random Replacement Replacement of Source-Language News

Monolingual Consumption Pattern. Fig. 8.3a displays the average ranking performance, in terms of NDCG@10, of the NNRs across the 14 target languages in xMIND, given increasing proportions of target-language news added to the training data. Although incorporating target-language news does mitigate some of the performance loss seen under ZS-XLT$_{\text{MONO}}$, our results show that when the target language comprises less than 30–40% of the training data, the recommenders perform worse than under ZS-XLT. We believe that this effect stems from the relatively short user histories (averaging 33 articles), where a small fraction of news in a different language can confuse the recommenders' user encoder. Moreover, while higher proportions of target-language data generally produce greater gains over ZS-XLT, the improvements differ by model and the amount of target-language news injected. For example, models such as NAML and CAUM tend to yield less accurate rankings when the share of target-language news is very low or very high, whereas MANNeR – featuring a non-parameterized user encoder – shows more stable performance across varying distributions of source and target-language articles. Examining representative results per language (Fig. 8.4), we find that FS-XLT particularly benefits low-resource languages, and languages not included in XLM-RoBERTa's pretraining. While NAML (Fig. 8.4b) and LSTUR rarely see significant gains from target-language injection, models such as TANR, and particularly MANNeR (Fig. 8.4a), reap substantial benefits from even a modest amount of target-language training data. In the case of MANNeR, this robustness could be attributed to its contrastive loss objective, which naturally pulls positive news (i.e., clicked articles) from the user history closer together in the embedding space, independent of language. For TANR, a plausible explanation is its joint recommendation-topic prediction objective, which places greater emphasis on topic representations that are largely language-agnostic. We note, in particular, that for languages such as KAT or THA, injecting any fraction of target-language news actually reduces



(a) Results for MANNeR (b) Results for NAML

Figure 8.4: Relative percentage difference in ranking performance (NDCG@10), under FS-XLT$_{\text{MONO}}$ compared to ZS-XLT$_{\text{MONO}}$, with a random replacement of English with target-language news in training.



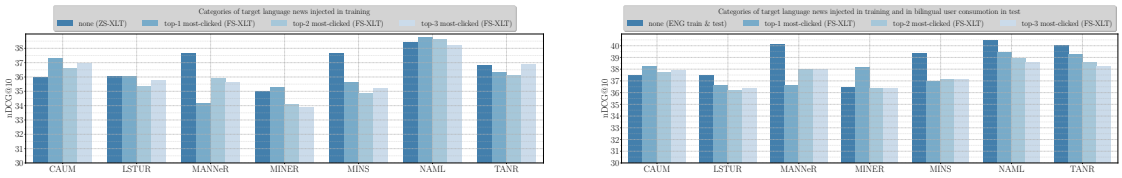
(a) Results for MANNeR (b) Results for NAML

Figure 8.5: Relative percentage difference in ranking performance (NDCG@10), under FS-XLT$_{\text{BILING}}$ compared to ZS-XLT$_{\text{BILING}}$, with a random replacement of English with target-language news.

ranking performance in NAML compared to the zero-shot setting. One possible explanation is that the additive-attention-based user encoder struggles to capture language-specific patterns from the short user histories.

Bilingual Consumption Pattern. In contrast to the FS-XLT$_{\text{MONO}}$ setting, few-shot target-language injection appears less effective under bilingual news consumption (i.e., FS-XLT$_{\text{BILING}}$). Fig. 8.3b plots the ranking performance (NDCG@10) of various NNRs in FS-XLT$_{\text{BILING}}$ compared to their respective ZS-XLT$_{\text{BILING}}$ baselines, for different ratios of target-language news injected during training. Most NNRs match or surpass their ZS-XLT$_{\text{BILING}}$ performance, indicating that limited exposure to target-language samples can be beneficial. However, NAML – generally the top-performing NNR in our evaluation (when using XLM-RoBERTa as the base model) – performs worse than the version trained solely on English. As seen in Fig. 8.5b, NAML only benefits from FS-XLT$_{\text{BILING}}$ for languages where it already suffers the greatest losses under ZS-XLT$_{\text{BILING}}$ (e.g., CMN, JPN, HAT, and GRN, cf. Fig. 8.2a). Similar trends emerge for CAUM and MINER, whereas MANNeR (Fig. 8.5a) and MINS achieve more consistent improvements across all languages.

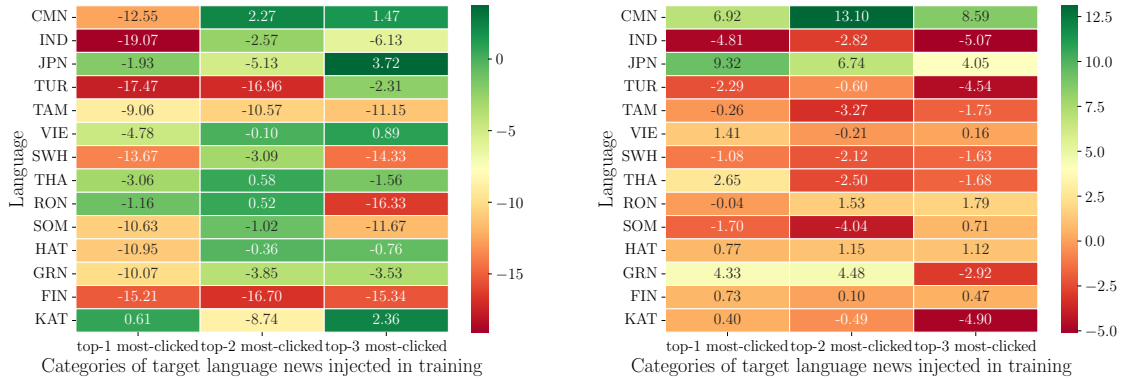
Overall, these performance variations, and the relatively modest gains from random-based few-shot target-language injection under bilingual consumption, emphasize the need for a deeper investigation into the specific factors influencing multilingual NNR performance. Such insights could guide the design of more robust user encoder architectures for multilingual user histories. Although we have performed a comprehensive analysis of encoder design in neural news recommenders in Chapter 5, our findings focus solely on monolingual settings in the English-only MIND dataset (Wu et al., 2020e), leaving open questions about multilingual settings.



(a) FS-XLT$_{\text{MONO}}$ vs ZS-XLT$_{\text{MONO}}$ results

(b) FS-XLT$_{\text{BILING}}$ results

Figure 8.6: FS-XLT ranking performance (NDCG@10) averaged over the 14 languages in xMIND, for various target-language injection ratios during training, and in the user's bilingual news consumption in test. We show the results for the *category-based replacement* strategy in FS-XLT.



(a) Results for MANNeR

(b) Results for NAML

Figure 8.7: Relative percentage difference in ranking performance (w.r.t. NDCG@10), under FS-XLT$_{\text{MONO}}$ compared to ZS-XLT$_{\text{MONO}}$, with a *category-based replacement* of target-language news in training.

8.6.2 Category-Based Replacement of Source-Language News

Next, we analyze the models' performance under FS-XLT with a category-based injection of target-language news during training. Specifically, we replace English news with target-language translations from a user's top-k most-clicked categories.

Monolingual Consumption Pattern. Fig. 8.6a illustrates the average ranking performance across all 14 languages in xMIND for the benchmarked recommenders under FS-XLT$_{\text{MONO}}$, varying $k \in \{1, 2, 3\}$, and compares these results to the ZS-XLT$_{\text{MONO}}$ baseline (i.e., no injected target-language news during training). We find that for 5 out of 7 models, category-based injection offers small to moderate gains over ZS-XLT$_{\text{MONO}}$. However, MANNeR and MINS experience significant performance drops under category-based injection. For the remaining models, injecting target-language news from only the single most-clicked category generally yields the highest relative improvement over FS-XLT$_{\text{MONO}}$, ranging from 0.31% for LSTUR to 3.64% for CAUM. Injecting more categories (i.e., $k = 2$ or $k = 3$) yields limited additional gains (e.g., CAUM, NAML) or even leads to performance decline (e.g., LSTUR, MINER). TANR is an exception, benefiting only when all top 3 most-clicked categories per user are replaced.

Fig. 8.7 compares category-based FS-XLT$_{\text{MONO}}$ to ZS-XLT$_{\text{MONO}}$ for NAML and MANNeR across all 14 languages in xMIND. For NAML (Fig. 8.7b), category-based injection tends to harm performance on higher-resource languages (e.g., IND, TUR, SWH) – regardless of the value of k – aligning with the behavior observed for MANNeR and MINS, and contrasting the ZS-XLT$_{\text{MONO}}$ results of other models. Nevertheless, category-based injection does improve performance for non-Latin scripts (e.g., CMN,

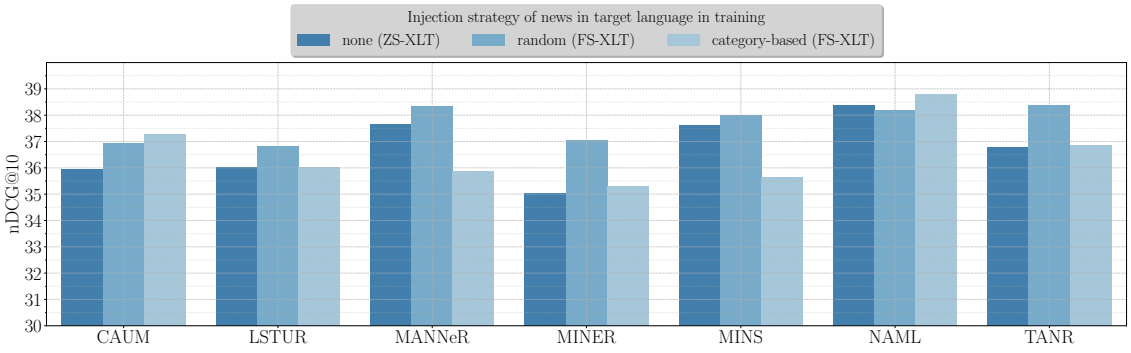


Figure 8.8: ZS-XLT$_{\text{MONO}}$ versus FS-XLT$_{\text{MONO}}$ ranking performance (NDCG@10) averaged across the 14 languages in xMIND. For the FS-XLT$_{\text{MONO}}$ setups, we consider both a *random* and a *category-based* replacement strategy. We report the performance with the best-performing percentage (for random replacement) or number of categories (for category-based replacement) of target-language news injected per model.

JPN) and low-resource languages (e.g., HAT, GRN) that were not seen during the pretraining of the base mPLM. This pattern holds across most models except MANNeR (Fig. 8.7a). However, we find that the optimal k value varies per language for each recommender.

Bilingual Consumption Pattern. Finally, we evaluate the recommenders under FS-XLT$_{\text{BILING}}$ with category-based replacement (Fig. 8.6b), averaging results across all languages and comparing them to the fully English (monolingual) training and testing setting (i.e., no target-language injection and monolingual history). Most models under bilingual consumption perform worse in terms of ranking compared to the monolingual English setting, despite target-language injection during training. However, CAUM and MINER show notable benefits from category-based target-language injection even in a bilingual news consumption context. These findings underscore the need for more advanced multilingual news encoders and user modeling approaches tailored to multilingual reading histories.

8.6.3 Random Versus Category-Based Replacement of Source-Language News

Lastly, we investigate whether category-based injection of target-language news during training offers any advantages over random-based injection. Fig. 8.8 compares the ranking performance of the NNRs in both ZS-XLT$_{\text{MONO}}$ and FS-XLT$_{\text{MONO}}$ conditions. For FS-XLT$_{\text{MONO}}$, we select each model's best-performing injection approach (i.e., the proportion of randomly replaced news or the number of replaced categories) as determined by the highest NDCG@10. Our findings show that random replacement outperforms category-based replacement for 5 out of 7 models; only CAUM and NAML see better results when the replaced news belongs to a user's most-clicked category.

To better interpret these differences, we examine the proportion of target-language news injected by each strategy. On average, a user's history consists of 33.05 clicked news items across 7.12 unique categories. The top most-clicked category contains an average of 12.79 news items, while the second and third most-clicked categories include 5.83 and 6.85 items per user, respectively.[19] The random replacement strategy's best-performing injection rates range between 40% (for MINER) and 80% (for TANR). In contrast, injecting only the top most-clicked category yields a 38.7% replacement rate – below the lowest best-performing random replacement ratio. As shown in Fig. 8.3 and discussed in Section 8.6.1, injecting less than 40% of target-language news offers little to no improvement over ZS-XLT. Increasing injection of news belonging to the top two (56.33%) or three (77.07%) categories brings the replacement rate closer to that achieved by random replacement (40% - 80% as noted

[19]We only report statistics for the training set, as those for validation and test sets are nearly identical.

above), but performance remains lower. We hypothesize that inclusion of more diverse categories in the target language constitutes a training signal that better generalizes to unseen topics.

In conclusion, randomly injecting target-language news during training generally delivers better FS-XLT performance, irrespective of the news topic. Nonetheless, further research is needed to develop effective cross-lingual transfer encoders that can narrow the performance gap between source (typically English) and target languages.

8.7 Further Analysis of Cross-lingual Transfer in Neural News Recommendation

To gain deeper insight into cross-lingual transfer in news recommendation, we conduct two additional analyses. First, we test whether more advanced fine-tuning strategies for the news encoder's underlying mPLM improve ZS-XLT and FS-XLT compared to simply unfreezing the last four layers. Second, we examine the degree of representation alignment between the source language (English) and the target languages within the mPLM during news encoding. Due to the high computational cost, we restrict these experiments to ZS-XLT and FS-XLT settings with *monolingual news consumption patterns* (e.g., user reading histories are exclusively in English or exclusively in the target language). We further narrow the scope to two of the best-performing NNRs identified in our earlier evaluations (cf. Section 8.5 and Section 8.6): NAML, representative of models trained with pointwise classification objectives, and MANNeR, trained with a supervised contrastive loss.

8.7.1 Fine-Tuning Strategies

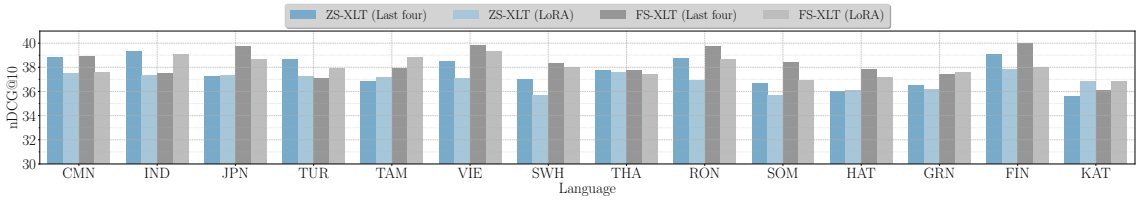
In the previous experiments, we updated only the parameters of the last four layers of the news encoder's underlying mPLM on the news recommendation task, a strategy which we denote *Last four*. We now compare this against a parameter-efficient fine-tuning technique. Specifically, we insert low-rank adapters (LoRAs, Hu et al., 2022), $\Delta W_{i=1}^{W}$, into the linear layers W of the mPLM. Only the LoRA parameters are updated during training, while all other mPLM parameters remain frozen. The goal is to examine whether retaining most of the pretrained knowledge while updating only a small number of additional weights can improve cross-lingual performance, while also significantly reducing the number of parameters that need to be fine-tuned.

Experimental Setup. We conduct ablation experiments using XLM-RoBERTa (Conneau et al., 2020) as the underlying mPLM in the news encoder of the recommenders. For LoRA, we tune its main hyperparameters, the dimension (rank r) and the scaling factor alpha (α) of the low-rank matrices, and the learning rate, independently for both evaluated NNRs. We perform a grid search with the learning rate in the range $[5e^{-5}, 1e^{-4}, 2e^{-4}, 5e^{-4}]$, r in $[16, 32]$, and α in $[32, 64]$. We train all models with the optimized learning rate of $5e^{-5}$ using LoRA with rank $r = 16$, a scaling factor alpha $\alpha = 32$, and LoRA dropout of 0.05 in all linear layers of the mPLM. We set all other model-specific hyperparameters to the optimal values reported in § 8.4.2 and Appendix F.2. In $\texttt{FS-XLT}_{\texttt{MONO}}$, we adopt the *random replacement* strategy with the best-performing target-language injection ratio per model, namely 60% for NAML, and 70% for MANNeR. We repeat each experiment three times, and report averages and standard deviations for the standard metrics: AUC, MRR, NDCG@10.

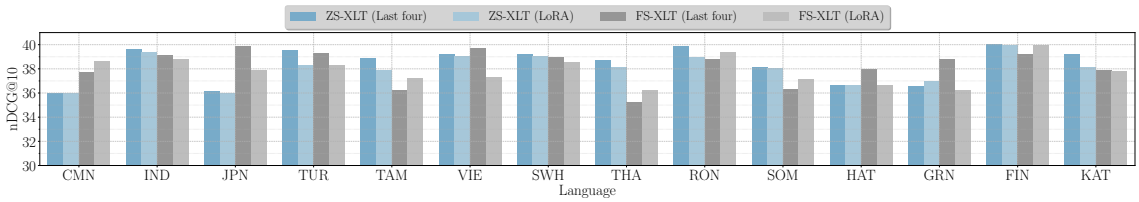
Results and Discussion. Table 8.7 compares the performance of the NNRs in $\texttt{ZS-XLT}_{\texttt{MONO}}$ under the two fine-tuning strategies. Using LoRAs instead of fine-tuning the last four layers of the mPLM reduces the number of trainable parameters in the NNRs by 98%. More importantly, NAML and MANNeR trained with LoRA achieve considerably better classification performance and comparable ranking performance relative to the *Last four* training variant. For MANNeR in particular, LoRA significantly narrows the ranking performance gap between English and target languages, reducing the relative drop in MRR from 7.15 to 1.90, and in NDCG@10 from 6.28 to 1.88. Although it may

Model	Fine-tuning	# Parameters (M)		AUC			MRR			NDCG@10		
		Trainable	Total	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ
MANNeR	Last four	222	279	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	0.00	$35.58_{\pm 0.31}$	$33.03_{\pm 0.54}$	-7.15	$40.17_{\pm 0.21}$	$37.64_{\pm 0.44}$	-6.28
	LoRA	4.3	282	$62.44_{\pm 3.18}$	$61.27_{\pm 2.31}$	-1.87	$33.11_{\pm 1.55}$	$32.48_{\pm 1.18}$	-1.90	$37.62_{\pm 1.61}$	$36.91_{\pm 1.21}$	-1.88
NAML	Last four	227	280	$52.85_{\pm 2.27}$	$52.49_{\pm 2.60}$	-0.68	$35.98_{\pm 0.44}$	$33.98_{\pm 0.95}$	-5.56	$40.43_{\pm 0.39}$	$38.38_{\pm 1.02}$	-5.06
	LoRA	5.6	283	$58.66_{\pm 0.82}$	$55.16_{\pm 1.83}$	-5.96	$35.77_{\pm 0.77}$	$33.49_{\pm 0.98}$	-6.36	$40.28_{\pm 0.58}$	$38.01_{\pm 0.97}$	-5.63

Table 8.7: ZS-XLT$_{\text{MONO}}$ recommendation performance with different fine-tuning strategies for NAML and MANNeR. For each model, we report the number of model parameters (in millions), and the performance (i) on the English MIND dataset (denoted ENG), (ii) averaged across all 14 target languages in xMIND (denoted AVG), and (iii) the relative percentage difference between average ZS-XLT$_{\text{MONO}}$ and ENG performance (%Δ). We report averages and standard deviations across three runs.



(a) Results for MANNeR



(b) Results for NAML

Figure 8.9: ZS-XLT$_{\text{MONO}}$ and FS-XLT$_{\text{MONO}}$ ranking performance (NDCG@10), across the 14 languages of xMIND, for different fine-tuning strategies of the mPLM (XLM-RoBERTa) in the recommenders' news encoder. In FS-XLT$_{\text{MONO}}$, we report the results for the *random replacement* strategy with the best-performing target-language injection ratio per model.

appear counterintuitive that LoRA improves classification while *Last four* performs better on ranking, one possible explanation lies in their different adaptation strategies. LoRA freezes the original mPLM weights and learns a low-rank update, which provides stronger regularization and preserves much of the pretrained representation, thereby reducing overfitting and often improving global classification calibration (Hu et al., 2022). Ranking metrics, however, depend on fine-grained local ordering of embeddings. Full or partial fine-tuning (i.e., *Last four*) can rotate many singular directions of the weight matrices to optimize relative ordering, whereas low-rank updates produce different structural updates that may be insufficient to induce the high-dimensional shifts needed for top-k ordering (Shuttleworth et al., 2024). More broadly, dense retrieval and ranking require high capacity and more aggressive feature learning to separate subtle semantic differences (Karpukhin et al., 2020). Since LoRA updates fewer effective degrees of freedom, it may underfit pairwise signals while performing well on per-sample label prediction (classification).

Figs. 8.9a and 8.9b break down the ZS-XLT$_{\text{MONO}}$ and FS-XLT$_{\text{MONO}}$ ranking performance (NDCG@10) across the 14 target languages of xMIND, for MANNeR, and NAML, respectively. The ZS-XLT$_{\text{MONO}}$ results reinforce the findings from Table 8.7: ranking performance either slightly deteriorates or remains stable across all languages when reducing the number of trainable parameters via LoRA fine-

tuning. In contrast, in FS-XLT$_{\text{MONO}}$, LoRA provides modest gains over *Last four* for certain languages, for example IND, TUR, and KAT in MANNeR, and CMN, TAM, and SOM in NAML. Although LoRA does not consistently yield large performance improvements, it offers a more parameter-efficient and composable alternative to partial or full fine-tuning. Given the prohibitive cost of fine-tuning large mPLMs, particularly in low-resource settings, LoRA and other lightweight methods such as adapter layers or language specific-adapters (Pfeiffer et al., 2020; Poth et al., 2023) warrant further exploration to mitigate performance degradation in cross-lingual transfer in news recommendation.

8.7.2 Representation Alignment Across Languages

We next examine the degree of alignment between source-language (English) and target-language representations of news articles produced by the mPLM before and after fine-tuning.

Experimental Setup. We use XLM-RoBERTa (Conneau et al., 2020) as the underlying mPLM for both recommenders, taking the output representation of the sequence start token $[CLS]$ as the news embedding. We consider the mPLM in two states: (i) out-of-the-box, i.e., without any task-specific fine-tuning, and (ii) after fine-tuning it in an FS-XLT setup with varying target-language injection ratios during NNR training. To isolate the effects of the mPLM itself and avoid confounding contributions from other encoder components (e.g., category or entity encoders), we only analyze the fine-tuned mPLM rather than the full news encoder. Concretely, we encode the titles of articles from the MIND test set (English) and xMIND test set (target language), both before and after fine-tuning.

Alignment Evaluation. We assess representation alignment along three dimensions: (i) pairwise alignment, (ii) subspace alignment, and (iii) representation similarity.

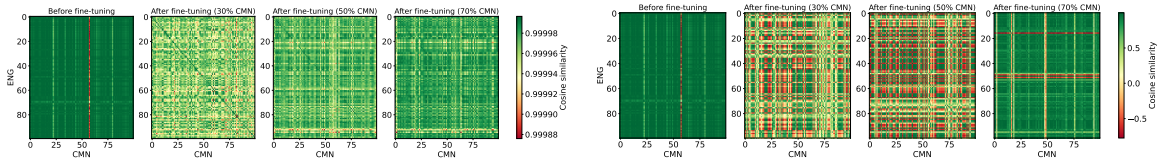
Pairwise Alignment. We measure pairwise alignment at the article-level using cosine similarity between English and target-language embeddings. Cosine similarity indicates whether representations of the same article in two languages become more similar after fine-tuning the underlying mPLM. However, it is sensitive to absolute shifts and orientation changes in the embedding space.

Subspace Alignment. To assess whether the embedding spaces of the source and target languages span similar directions, we measure subspace alignment using principal angles and Grassmann distance. Formally, given two k-dimensional subspaces $U, V \subset \mathbb{R}^d$, the principle angles $\theta = [\theta_1, \theta_2, \ldots, \theta_k] \in [0, \frac{\pi}{2}]$ are defined recursively as:

$$\begin{gathered} \cos\theta_k = \max_{\substack{\mathbf{u_k}\in U \\ \mathbf{v_k}\in V}} \mathbf{u_k}^T\mathbf{v_k}, \text{ subject to} \\ \mathbf{u_k}^T\mathbf{u_k} = 1, \mathbf{v_k}^T\mathbf{v_k} = 1, \mathbf{u_k}^T\mathbf{u_i} = 0, \mathbf{v_k}^T\mathbf{v_i} = 0, (i = 1, ..., k-1) \end{gathered} \tag{8.1}$$

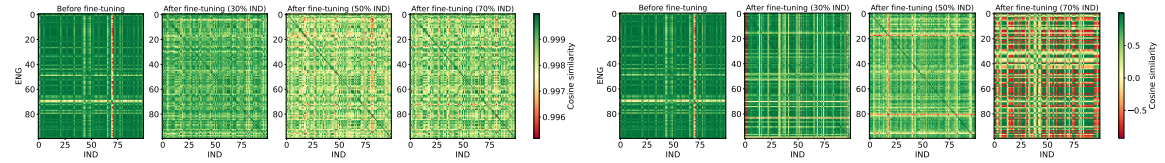
The first principal angle θ_1 is the smallest angle between any pair of unit vectors from U and V (i.e., the most aligned directions). Subsequent angles measure alignment between the next closest directions, under orthogonality constraints (Hamm and Lee, 2008). If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ denote orthonormal bases of U and V, then the cosines of the principal angles are given by the singular values of $\mathbf{Y}_1^T\mathbf{Y}_2$. Small angles indicate strong overlap between subspaces, while large angles reflect near-orthogonality. The Grassmann (geodesic) distance provides a scalar summary (Hamm and Lee, 2008): $d_{Grass}(U, V) = \left(\sum_{i=1}^{k} \theta_i^2\right)^{\frac{1}{2}}$. Intuitively, this distance measures the global misalignment between two subspaces. In our experiments, we compare the top-k ($k = 50$) PCA subspaces of the source- and target-language embeddings, reporting both the distribution of principal angles and the Grassmann distance before and after fine-tuning.

Representation Similarity. Finally, we evaluate structural similarity between source- and target-language embeddings using the Central Kernel Alignment (CKA) metric (Kornblith et al., 2019). Let $X, Y \in \mathbb{R}$ denote two sets of embeddings of the same n instances, with linear kernels (e.g., similarity matrices) $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ and $\mathbf{L} = \mathbf{Y}\mathbf{Y}^T$. CKA is defined as the normalized Hilbert-Schmidt



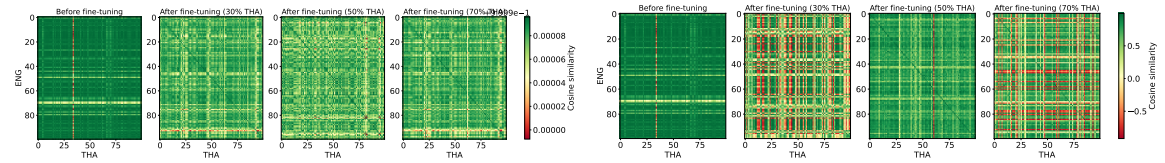
(a) MANNeR: CMN

(b) NAML: CMN



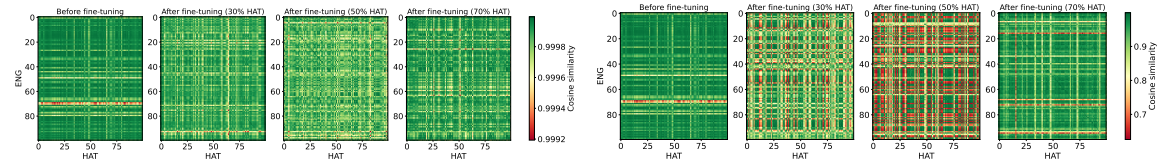
(c) MANNeR: IND

(d) NAML: IND



(e) MANNeR: THA

(f) NAML: THA



(g) MANNeR: HAT

(h) NAML: HAT

Figure 8.10: Cosine similarity between the embeddings of English and target-language news articles in the test set, before and after fine-tuning the mPLM (XLM-RoBERTa) in the NNRs' news encoder, across selected languages of xMIND. We report results for $\texttt{FS-XLT}_{\texttt{MONO}}$, using the *random replacement* strategy and three target-language injection ratios during training (e.g., 30%, 50%, and 70%).

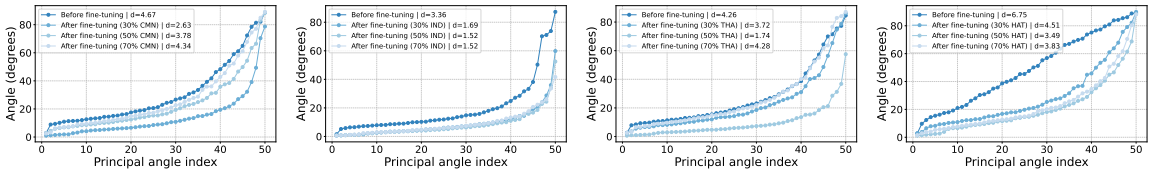
Independence Criterion (HSIC Gretton et al., 2005):

$$CKA(\mathbf{K}, \mathbf{L}) = \frac{HSIC(\mathbf{K}, \mathbf{L})}{\sqrt{HSIC(\mathbf{KK})HSIC(\mathbf{LL})}} \tag{8.2}$$

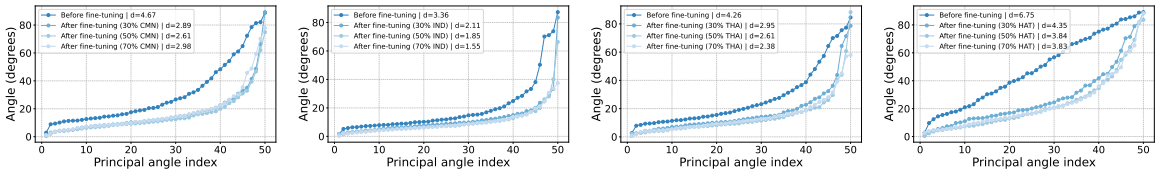
where $HSIC(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2} tr(\mathbf{KHLH})$, and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix. CKA values lie in $[0, 1]$, with 1 indicating identical relational structure, and 0 denoting unrelated representations. In our context, higher CKA scores between source- and target-language embeddings after fine-tuning suggest that the relational geometry across languages becomes more similar.

Results and Discussion. Fig. 8.10 shows heatmaps of cosine similarities between English and selected target languages from xMIND, before and after fine-tuning, for both MANNeR (left-hand side) and NAML (right-hand side).[20] The diagonal of each heatmap represents the similarity of the same article across English and the target language. In the pre-fine-tuning heatmaps, the diagonal is fuzzy with pairwise matches not being clearly distinguishable from off-diagonal entries (i.e., non-matches). This suggests that the embedding spaces are poorly aligned, as confirmed by many strong off-diagonal similarities. After fine-tuning, the diagonals generally become sharper, with lower off-diagonal cosine similarities (lighter colors). This indicates improved cross-lingual alignment, with English and the target-language counterparts of the same article moving closer in the embedding

[20]For brevity, we illustrate only representative examples of target languages from xMIND in this section. However, our analysis in the discussion of results covers all 14 target languages.



(a) MANNeR: CMN (b) MANNeR: IND (c) MANNeR: THA (d) MANNeR: HAT



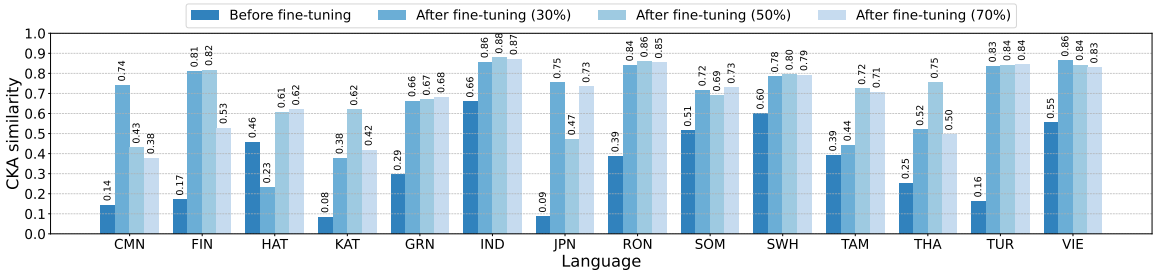
(e) NAML: CMN (f) NAML: IND (g) NAML: THA (h) NAML: HAT

Figure 8.11: Distribution of principal angles for the top $k = 50$ PCA subspaces, and Grassmann distance, for the embeddings of the English and the target-language news articles in the test set, before and after fine-tuning the mPLM (XLM-RoBERTa) in the recommenders' news encoder, across selected languages of xMIND. We report results for $\text{FS-XLT}_{\text{MONO}}$, using the *random replacement* strategy and three target-language injection ratios during training (e.g., 30%, 50%, and 70%).

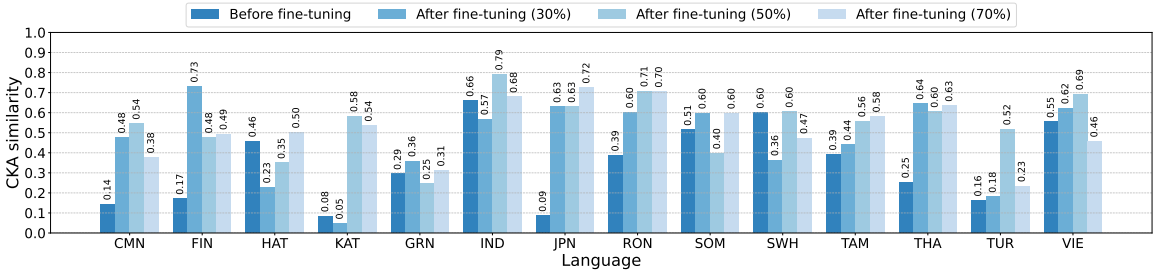
space. Still, consistent with Section 8.6, the benefits of different target-language injection ratios vary by language. Higher-resource languages such as CMN, FIN, JPN, SWH, or VIE achieve good alignment with only 30% to 50% target-language injection during fine-tuning. In contrast, languages such as HAT, KAT, or GRN require higher proportions of target-language training examples to improve cross-lingual alignment. These findings partially mirror the downstream raking performance trends discussed in § 8.6.1 (see Fig. 8.4): greater alignment seems to correspond to better downstream performance, though the optimal target-language injection ratios may vary.

We next turn to global alignment between the English and target-language embedding spaces. Fig. 8.11 shows the distribution of principal angles and corresponding Grassmann distances before and after fine-tuning. These metrics confirm our earlier observations: source- and target-language embedding spaces become more aligned after fine-tuning, as indicated by smaller angles and lower Grassmann distances. Before fine-tuning, larger principal angles indicate that source and target languages occupy different directions in the embedding space. After fine-tuning, these angles shrink, showing that the spaces move closer together. We observe the largest drop for IND, where the maximum principal angle decreases by over 40° with a 70% target-language injection ratio for both models. However, even after fine-tuning, low-resource languages remain less aligned with English than higher-resource languages, as reflected in their relatively high Grassmann distances. We also observe that different languages benefit from different target-language injection ratios. Notably, for MANNeR, certain ratios fail to improve, or even hurt, alignment: 70% for FIN, 50% for JPN, and 70% for THA. Interestingly, this contrasts with the downstream performance results in § 8.6.1 (see Fig. 8.4), where performance was less affected. This discrepancy may arise from differences in the training regimes of the two models, though further research is needed to better understand cross-lingual alignment in NNRs.

Finally, we examine representation similarity between source and target languages. Fig. 8.12 reports CKA scores for MANNeR and NAML across the 14 xMIND languages, before and after fine-tuning. Results again corroborate previous findings: fine-tuning generally increases CKA, reflecting more similar relational structures between source and target-language embeddings. We highlight a few noteworthy observations. For some languages, specific injection ratios decrease similarity: e.g., 30% for HAT (both models), 30% for KAT, IND, or SWH (NAML), and 50% for GRN (NAML). Additionally, for MANNeR, CKA scores for some languages remain similar across injection ratios (e.g., 30%, 50%, 70%), implying that fewer target-language examples may suffice. This is particularly encouraging



(a) Results for MANNeR



(b) Results for NAML

Figure 8.12: CKA similarity between the embeddings of English and target-language news articles in the test set, before and after fine-tuning the mPLM (XLM-RoBERTa) in the NNRs' news encoder, across the 14 languages of xMIND. We report results for $\texttt{FS-XLT}_{\texttt{MONO}}$, using the *random replacement* strategy and three target-language injection ratios during training (e.g., 30%, 50%, and 70%).

for resource-poor languages such as HAT, GRN, or SOM, where training data is scarce. These results thus show that comparable representation similarity can be achieved with fewer samples.

Overall, our findings suggest that task-specific fine-tuning with even a limited number of target-language instances improves both local and global cross-lingual alignment. Nevertheless, the embedding spaces of low-resource languages remain only weakly aligned with English, even after few-shot target-language fine-tuning. This highlights the need for further research into the factors driving cross-lingual NNR performance, and into strategies for improving representation alignment to better serve speakers of underrepresented languages. A promising direction is to leverage multilingual LLMs to reduce the semantic gap across languages. As discussed in § 8.2.2, LLMs can help mitigate sparsity by generating metadata (e.g., summaries, categories) in a pivot language from item descriptions in another language, or serve as a bridge between languages by rephrasing, translating, or contextualizing user histories across languages. Beyond improving semantic alignment, future work should also explore how LLMs can capture language-dependent preferences that extend beyond literal overlaps - for instance, recognizing that a user reading "fotbal" articles in Romanian might also be interested in "soccer" content in English. Since cross-lingual transfer effectively constitutes a cold-start scenario for the target language, future work could adapt methods from cold-start recommendation, such as exploiting side information or identifying transferable behavior patterns, within LLM-driven frameworks to build more accurate and equitable cross-lingual recommenders.

8.8 Dataset Analysis

We finally investigate the quality of the translations in xMIND. Concretely, we (i) estimate the translation quality and (ii) investigate the robustness of the NNRs to different translations of the source news from MIND (Wu et al., 2020e). To this end, we use xMINDsmall and additionally translate its training and development portions using Google (Neural Machine) Translation (GNMT) (Wu et al., 2016), a commercial MT system that supports all xMIND languages. GNMT has been

	Language	CMN	IND	JPN	TUR	TAM	VIE	SWH	THA	RON	SOM	HAT	GRN	FIN	KAT
Intelligibility	NLLB	0.36	0.37	0.39	0.33	0.02	0.35	0.45	0.46	0.51	0.19	0.20	0.10	0.21	0.92
	GNMT	0.30	-0.01	0.09	0.12	0.34	0.13	0.33	-0.11	0.26	-0.08	-0.06	-0.05	-0.21	0.00
Fidelity	NLLB	0.55	0.42	0.36	0.48	0.51	0.25	0.20	0.40	0.55	0.34	0.17	0.30	0.34	0.61
	GNMT	0.28	0.21	0.20	0.39	0.10	0.14	0.19	0.02	0.37	-0.17	0.01	0.03	0.09	0.12
Pairwise Comparison		0.01	0.22	0.19	0.33	0.26	0.16	0.22	0.01	0.30	0.17	0.09	0.16	0.28	0.65

Figure 8.13: Annotator agreement (Krippendorf's alpha), per language, for three types of translation aspects in the annotation task.

shown to outperform NLLB on translation from English to various low-resource languages (Team et al., 2022). [21] [22]

8.8.1 Manual Quality Estimation of Translation

Given the size of the xMINDsmall dataset and our (limited) annotation budget, it was infeasible to manually post-edit the translations of the news in the test portion of xMIND. We therefore resorted to conducting an annotation task to estimate the quality of the translations. To this end, we sample 50 news from the development portion of the MIND dataset (i.e., the portion used as test set in all our experiments), according to the (i) the distribution of categories in the dataset and (ii) the distribution of the total length of the news (i.e., composed of title and abstract). This way, we ensure that the sampled instances are representative of the full dataset.

We carry out the annotations using the Potato annotation tool (Pei et al., 2022). Two annotators judged the quality of the NLLB and GNMT translations for each language.[23] The task comprised of a total of five questions, targeting three aspects of the translations: *intelligibility*, *fidelity*, and *pairwise comparison* between the NLLB and GNMT translations. The first two questions were repeatedly asked independently for the NLLB and GNMT translations. The annotators answered the following questions: *(1-2) Is the translation acceptable?* – binary answer; *(3-4) To which extent is the information from the original text accurately retained in the translation?* – 5-point Likert-scale answers, ranging from "Not at all" to "Completely"; *(5) Which translation is better?* – categorical answer with three options, namely "Translation A", "Translation B", or "They are comparably good". In order to remove any position bias in all questions, we randomized the source of translations A and B shown to the annotator, such that 50% of the time translation A stemmed from NLLB, and the remaining 50% from GNMT. Overall, across most of the target languages, we observed higher annotators agreement (Krippendorf's alpha (Krippendorff, 2013)) for the NLLB translations for the first two questions, as shown in Fig. 8.13, than for GNMT, where we observe little to no agreement between the annotators.

For over half of the target languages in xMIND, the annotators deemed the NLLB translations to be intelligible in at least 60% of the cases (see Fig. 8.14a). Similarly, we find that our translations retain the information of the original texts, at least partially, in the majority of cases, as illustrated in Fig. 8.14b. Notably, the NLLB-sourced translations are deemed more faithful to the original news than the GNMT-based ones particularly for low-resource languages such as TAM and GRN. This finding is corroborated by the results from the pairwise comparison (cf. Fig. 8.14c), which show that NLLB translations are judged to be overall better than their GNMT counterparts for these two

[21] GNMT is a proprietary system, hindering fair comparisons to open-source models due to the lack of transparency regarding its model architecture and training procedures.

[22] We translated the text with the Cloud Translation - Advanced (v3) API: https://cloud.google.com/translate/docs/overview, using Google Cloud research credits worth approximately $5,000.

[23] All annotators were native speakers of the target language and fluent in English. Most of the annotators were certified interpreters/translators of the target language.

languages. Nonetheless, across all languages and aspects of evaluation, translations obtained with the commercial GNMT are deemed generally of higher quality than those generated with the open-source NLLB.

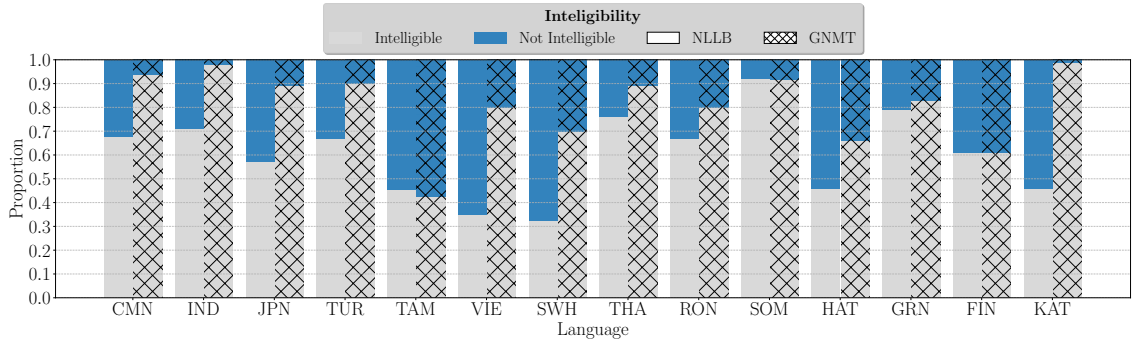
The annotators' feedback revealed several challenges that contributed to the generally low scores assigned to both kinds of translations. Firstly, we remark that often one part of the news (e.g., title) was perfectly translated, whereas the other portion (e.g., abstract) was not accurately depicted in the target language. Secondly, in few cases, the phrasing of the news title was hard to comprehend even in English, impeding translation. Lastly, we note that given the US origins of the news articles from the MIND dataset (Wu et al., 2020e), many of the topics discussed in the news are not usually encountered in some of the target languages or they pertain solely to the US (e.g., sports news). In such cases, we observed that the MT systems performed particularly poorly. A closer look at the translation errors reveals that, in terms of intelligibility, both NLLB and GNMT generate worse translations for news in categories such as entertainment (e.g., for languages such as VIE, TUR, JPN), movies, music, or television (e.g., particularly for lower-resource languages TAM and KAT). Such errors can be explained by the fact that these categories of news tend to contain terms that exhibit more idiomaticity (e.g., especially in movie titles), which is well-documented source of trouble for MT (Dankers et al., 2022). Similar patterns emerge when analyzing the news categories on which the fidelity of the translations is lower. However, our results indicate that there is not a particular category on which one of the MT systems is better than the other (according to our annotators). Lastly, we observe that translations of shorter texts, obtained with both NLLB and GNMT, are of higher quality than those of longer ones. This can be explained by the fact that, at least NLLB, has been trained on pairs of shorter documents (Team et al., 2022).

8.8.2 Robustness of NNRs to Translation Quality

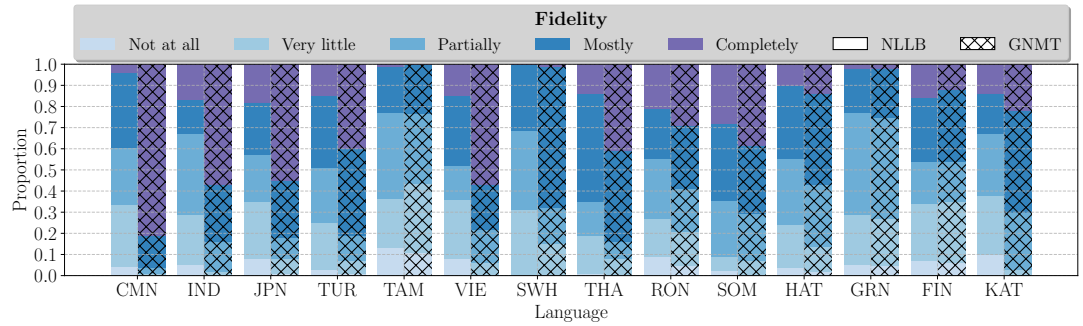
We next investigate the robustness of the best-performing NNRs from our previous experiments, namely NAML and MANNeR, to translations obtained with different MT systems. To this end, we re-run the previous experiments for the version of the dataset translated with GNMT and compare the results against the ones obtained using the NLLB-translated xMINDsmall.

Fig. 8.15 shows the ranking performance (w.r.t. NDCG@10) for both models and MT systems under ZS-XLT$_{\text{MONO}}$ (i.e., training on the English MIND and evaluation on the target language). The performance of both NNRs seems largely unaffected by the provenance of the translations. The differences are insignificant according to an independent samples T-test (for a p-value of 0.05), with the exception of MANNeR's performance on IND and KAT, where GNMT test translations lead to better ranked recommendations. This is important as it indicates that, although the GNMT translations were judged to be better on average than the ones produced by NLLB, the NNRs appear to be robust to differences in translation quality.

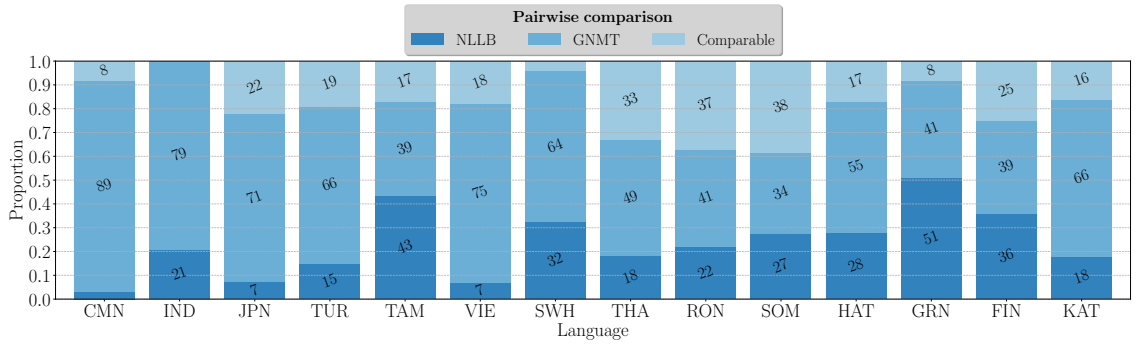
We further check this hypothesis in the ZS-XLT$_{\text{BILING}}$ setting. The corresponding results from Fig. 8.16a demonstrate small differences in performance depending on the MT system used. However, the differences are again not statistically significant according to the same independent T-test. We observe similar patterns and no statistical significance in both FS-XLT settings (e.g., Fig. 8.16b illustrates the ranking performance for the FS-XLT$_{\text{MONO}}$ case). Overall, these findings indicate that, despite the infeasibility of manual post-editing of the test set translations in xMIND, the quality of the translations obtained with the open-source NLLB (i) is on par with those generated by a state-of-the-art commercial MT system, and (ii) has no significant effect on the NNR's recommendation performance.



(a) Intelligibility of translations



(b) Fidelity of translations



(c) Pairwise comparison of translations

Figure 8.14: Annotation task results for each type of translation aspect.

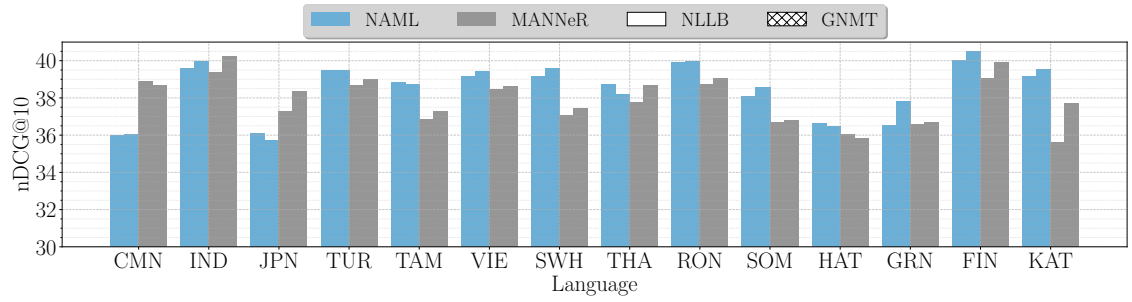
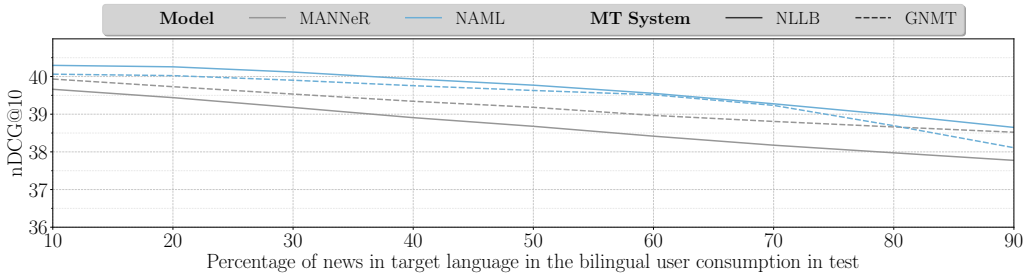
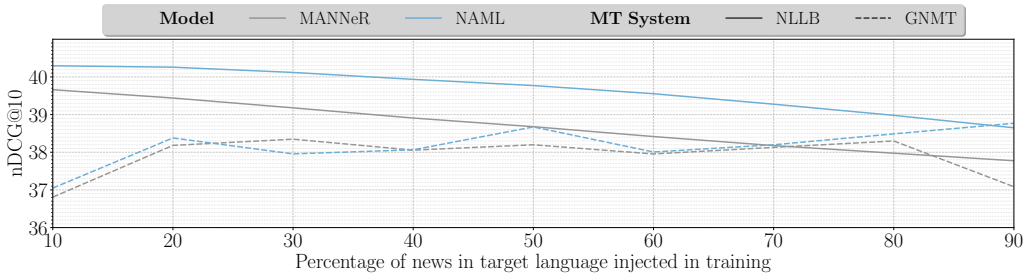


Figure 8.15: ZS-XLT$_{\text{MONO}}$ ranking performance (NDCG@10) in terms of MT system for NAML and MANNeR.



(a) ZS-XLT$_{\text{BILING}}$ results



(b) FS-XLT$_{\text{MONO}}$ results

Figure 8.16: Ranking performance (NDCG@10) in terms of MT system for NAML and MANNeR.

8.9 Conclusion

The ever-growing linguistic diversity of online news readers has yet to be fully addressed in news recommendation research, which remains predominantly focused on resource-rich languages – particularly English – and on monolingual consumption. In this Chapter, we introduced xMIND, an open-access, multilingual news recommendation dataset derived from the English-only MIND dataset via machine translation. xMIND encompasses 14 linguistically and geographically diverse languages. We use xMIND to benchmark several state-of-the-art content-based NNRs under zero-shot and few-shot cross-lingual transfer settings, exploring both monolingual and bilingual consumption scenarios. We consider two strategies for injecting target-language data during training, as well as constructing bilingual news consumption patterns: randomly replacing a portion of English news with their translations from xMIND, and replacing only English news corresponding to a user's top-clicked categories. Our findings reveal that existing NNRs experience substantial performance losses under ZS-XLT. Moreover, introducing target-language data during FS-XLT yields only modest gains, especially in bilingual news consumption contexts. Interestingly, random replacement generally outperforms category-based replacement, producing superior ranking performance in most NNRs. Additionally, our analysis of representation alignment between English and target languages within the news encoder's language model show that few-shot target-language injection during training primarily benefits high-resource languages. In contrast, low-resource languages remain weakly aligned with English and require substantially more target-language examples to achieve meaningful cross-lingual alignment. We believe xMIND serves as a valuable resource for the news recommendation community, and hope that it will foster more research on multilingual and cross-lingual news recommendation, for speakers of both high- and low-resource languages.

Having established a suitable news recommendation dataset for the evaluation and development of multilingual and cross-lingual news recommender systems, the next Chapter focuses on adapting a massively multilingual SE to the news domain in order to enable more robust and inclusive news recommendation across languages.

CHAPTER 9

NEWS DOMAIN ADAPTATION OF SENTENCE ENCODERS

Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt.
(The limits of my language mean the limits of my world.)

LUDWIG WITTGENSTEIN
«LOGISCH-PHILOSOPHISCHE ABHANDLUNG »

[*] Rapidly growing numbers of multilingual news consumers pose an increasing challenge to news recommender systems in terms of providing customized recommendations. First, existing neural news recommenders, even when powered by multilingual language models, suffer substantial performance losses in zero-shot cross-lingual transfer, as we have showed in Chapter 8. Second, the current paradigm of fine-tuning the backbone LM of a neural recommender on task-specific data is computationally expensive and infeasible in few-shot recommendation and cold-start setups, where data are scarce or completely unavailable.

Adapting the underlying LM to specific languages and domains (i.e., injecting news-domain knowledge into a PLM in one or more languages) is a promising strategy for reducing linguistic barriers, promoting inclusivity in news recommendation, and ensuring that NNRs serve a diverse set of users and scenarios. This approach allows the LM not only to handle languages that were unseen during pretraining, but also to capture the linguistic style and nuances of news content without requiring additional fine-tuning on recommendation-specific data. Ultimately, the goal is to enable users to access relevant information without being constrained by language, making multilingual news recommendation more equitable and effective.

In this Chapter, we propose a news-adapted sentence encoder (NaSE), domain-specialized from a pretrained massively multilingual sentence encoder (**C4**). To this end, we compile and leverage PolyNews and PolyNewsParallel, two multilingual news-specific corpora. With the news-adapted multilingual SE in place, we test the effectiveness of (i.e., question the need for) supervised fine-tuning for news recommendation, and propose a simple and strong baseline based on (i) frozen NaSE embeddings and (ii) late click behavior fusion. We show that NaSE achieves state-of-the-art performance in ZS-XLT in true cold-start and few-shot news recommendation.

9.1 Introduction

News recommender systems constitute the primary instrument used by digital news platforms to cater to the individual information needs of readers. The ever increasing language diversity of online

[*]This Chapter is adapted from: **Andreea Iana**, Fabian David Schmidt, Goran Glavaš, and Heiko Paulheim. 2025. News without Borders: Domain Adaptation of Multilingual Sentence Embeddings for Cross-lingual News Recommendation. In *Advances in Information Retrieval. ECIR 2025*, pages 119–137. Springer Nature Switzerland.

users has given rise to new challenges for news recommenders. Recommender systems must not only produce suitable, balanced, and diverse recommendations for multilingual news consumers from a variety of language backgrounds, but should also accurately perform in cold-start scenarios, where news data, user click logs, or both, are missing.

On the one hand, recent advancements in pretrained (multilingual) language models, used as the backbone of NNRs, has allowed to extend neural news recommendation beyond monolingual settings (Wu et al., 2021b; Guo et al., 2023). In cross-lingual transfer (XLT), however, even NNRs based on multilingual LMs display drastic performance loss in target-language recommendation compared to their source-language (usually English) recommendation performance (cf. Chapter 8). Although accurate XLT is critical especially for resource-lean(er) languages with limited-to-no click behavior data, strong multilingual sentence encoders (SEs) – trained exactly to align sentence semantics across a large number of languages (including many low-resource ones) – have largely been left unexplored as news encoding backbones in NNRs.

On the other hand, current NNRs typically fine-tune the underlying LM on task-specific data. Not only is fine-tuning a time and resource-intensive process, it is, more critically, infeasible in many real-world scenarios, with: little-to-no news-click data (i) for news in the target language (for particularly low-resource ones) or (ii) for the specific user (the so-called cold-start problem) (Wu et al., 2024c). Even when using frozen embeddings, most NNRs strictly require in-domain data to learn meaningful user representations for prediction (i.e., to train their parameterized user encoders to aggregate the embeddings of consumed news).

In this context, multilingual sentence encoders, which align sentence-level semantics across a wide range of languages (Yang et al., 2019a; Artetxe and Schwenk, 2019; Yang et al., 2020b; Feng et al., 2022), hold the promise of reducing the performance gap in XLT for NNRs. Multilingual SEs, however, have not been trained for news encoding, i.e., they lack the news domain-specific knowledge, which is crucial for performance in XLT news recommendation and cold-start scenarios. Moreover, even if equipped with robustly domain-adapted LMs, the majority of NNRs would still require fine-tuning on click behavior data to learn the weights of their trainable user encoder modules.

Contributions. We address the above limitations and advance cross-lingual and cold-start news recommendation with the following contributions: **(1)** We compile PolyNews[1] and PolyNewsParallel[2], two multilingual news-specific corpora which can be used not only for domain-adaptation of existing LMs, but also for machine translation (MT). **(2)** We train a news-adapted multilingual sentence encoder (dubbed NaSE) by domain-specializing a general-purpose multilingual SE on PolyNews and PolyNewsParallel with denoising auto-encoding and MT objectives.[3] **(3)** Leveraging frozen NaSE embeddings and resorting to non-parameterized late click behavior fusion (cf. Chapter 4), we introduce a simple and strong recommendation technique that yields state-of-the-art performance in zero-shot cross-lingual transfer (ZS-XLT) recommendation in cold-start setups, as well as in few-shot recommendation. This challenges the established paradigm of fine-tuning LMs for news recommendation on click behavior data.

9.2 Related Work

News Recommendation. Personalized news recommenders attenuate the information overload for readers by generating suggestions customized to their preferences (Li and Wang, 2019; Wu et al., 2023). The existing NNRs have two drawbacks. First, the quality of embeddings produced by news encoders with vanilla multilingual LMs seems inadequate for XLT, with substantial performance losses for target languages, as we have shown in Chapter 8. Yet, the usage of multilingual SEs –

[1] https://huggingface.co/datasets/aiana94/polynews
[2] https://huggingface.co/datasets/aiana94/polynews-parallel
[3] https://huggingface.co/aiana94/NaSE

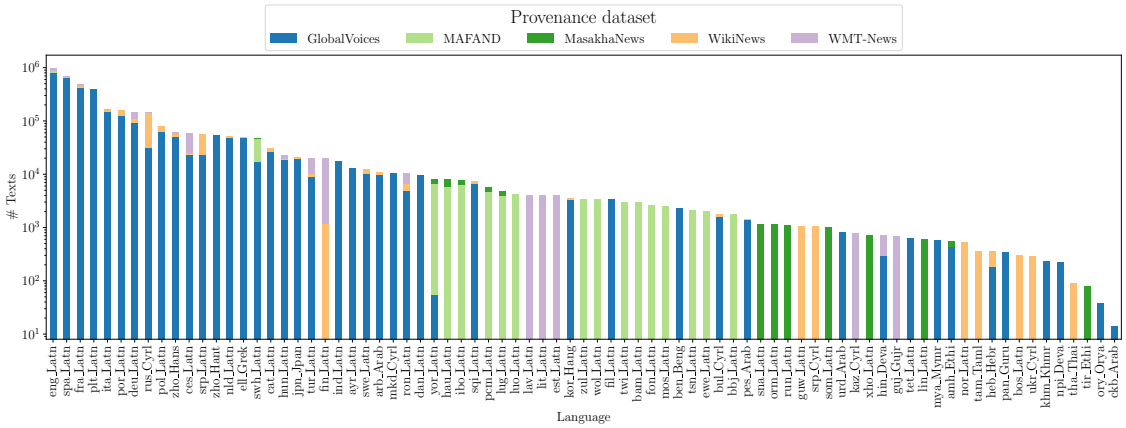


Figure 9.1: Distribution of texts (log-scale) in PolyNews across languages and provenance datasets.

precisely trained for cross-lingual semantic alignment across languages on parallel data – as backbone for news encoders is unexplored. Second, *fine-tuning* of the news encoder (i.e., its LM) is required, which is both (i) computationally expensive, as it updates the LM's hundreds of millions of parameters on large-scale click behavior datasets, and, more critically, (ii) infeasible in setups with limited or no click behavior data (e.g., with news in resource-lean languages or in cold-start setups, with no prior user behavior). In this Chapter, we address these limitations by adapting a general-purpose multilingual SE for the news domain. As the news encoder's backbone, this enables robust XLT for news recommendation, and supports setups where task-specific fine-tuning is not possible.

Domain-Adaptation of Language Models. The most common approach for injecting domain knowledge into LMs is pretraining on in-domain data with self-supervised objectives, e.g., Masked Language Modeling (Devlin et al., 2019) or SimCSE (Gao et al., 2021; Liu et al., 2021a; Wang et al., 2022d). Specializing to a particular domain is done either from scratch (Beltagy et al., 2019; Wu et al., 2020a; Lee et al., 2020b) or by adapting already pretrained LMs (Gururangan et al., 2020; Hung et al., 2022). Training from scratch is computationally intensive and requires large-scale domain-specific corpora, often difficult to obtain (Wang et al., 2022a; Hung et al., 2023). Adaptation instead is a lighter-weight alternative, as it starts from the already pretrained LM, requiring more moderate amounts of in-domain data. For SEs specifically, domain specialization is achieved by training for similarity or relevance estimation tasks with various self-supervised objectives (Liu and Yang, 2022) or via in-domain data generation (Wang et al., 2022a). Current work mainly derives domain-specific SEs from general-purpose (i.e., not sentence-specialized) pretrained LMs. In this Chapter, in contrast, we adapt an existing multilingual SE on in-domain data using denoising auto-encoding and machine translation objectives.

9.3 Domain-Specialized Corpora

A critical aspect of successful domain adaptation of multilingual LMs is the availability of high-quality training data. To this end, we first compile two large-scale multilingual corpora by leveraging existing news data collections.

PolyNews. We compile the multilingual non-parallel corpus `PolyNews` by combining news from the following five sources: (i) WikiNews (May 2024 dump)[4], (ii) the `train` split of MasakhaNews (Adelani et al., 2023), (iii) the `train` split of MAFAND (Adelani et al., 2022), (iv) WMT-News (`v2019`), and (v) GlobalVoices (`v2018q4`) (Tiedemann, 2012). We group together the articles from

[4] https://www.wikinews.org/

all five sources according to language-script combinations (some languages, e.g., Serbian, have multiple scripts). We preprocess the resulting corpus by removing exact duplicates, and news written in scripts not corresponding to the language of the sub-collection (e.g., Arabic texts in a French or English collection). We use GlotLID (Kargaran et al., 2023), a FastText-based model (Joulin et al., 2017) for language identification. In addition, we remove the $K\%$ shortest texts per language (in terms of character length), as these correspond to sequences of words that do not form meaningful sentences.[5] We determine K separately for each of the five news collections, as the quality of the texts heavily depends on their provenance. WikiNews, for example, contains many short (e.g., less than five words) and uninformative texts, which we consider noise.[6] Lastly, given the importance of avoiding text duplication in LM training (Lee et al., 2022), we perform MinHash (Broder, 1997) near de-duplication over all sentences of a language.[7] Finally, PolyNews contains 3,913,873 news texts, covering 77 languages and 19 scripts. Fig. 9.1 shows its distribution across languages and provenance. We further profile PolyNews in Appendix G.1.

PolyNewsParallel. We compile the translation corpus `PolyNewsParallel` from the parallel news collections MAFAND (Adelani et al., 2022), WMT-News, and GlobalVoices (Tiedemann, 2012). We use the same preprocessing pipeline as for PolyNews, and remove near-duplicated texts from the source language. Our final parallel corpus contains 5,386,846 texts over 833 language pairs, spanning 64 languages and 17 scripts.[8]

9.4 Methods for Domain Specialization

We obtain the news-adapted sentence encoder NaSE via sequence-to-sequence training on PolyNews and PolyNewsParallel of a massively multilingual sentence encoder. NaSE can be either (i) further fine-tuned downstream for news recommendation (i.e., on click behavior data) or (ii) directly used as a strong news encoder in cross-lingual news recommendation, without any fine-tuning (see Section 9.6).

9.4.1 Domain Adaptation

Our first training objective (DAE) is an adaption of the transformer-based sequential denoising auto-encoding (TSDAE) approach from (Wang et al., 2021a), which we use to specialize a pretrained multilingual SE for the news domain. TSDAE encodes a corrupt version of the input sentence, obtained by adding discrete noise (e.g., token deletion), and then learns to reconstruct the original sentence from the encoding of the noisy input. TSDAE can be formalized as follows:

$$\begin{aligned}\mathcal{J}_{TSDAE}(\Theta) &= \mathbb{E}_{x\sim D}\left[\log P_\Theta(x|\tilde{x})\right] = \mathbb{E}_{x\sim D}\left[\sum_{t=1}^{l}\log P_\Theta(x|\tilde{x})\right] \\ &= \mathbb{E}_{x\sim D}\left[\sum_{t=1}^{l}\log\frac{exp(h_t^T e_t)}{\sum_{i=1}^{|V|}exp(h_t^T e_i)}\right]\end{aligned} \tag{9.1}$$

where D is the training corpus, $x = x_1x_2...x_l$ the input sentence with l tokens, $\tilde{x}$ the corresponding corruption, e_t the sequence embedding of x_t, $|V|$ the vocabulary size, and h_t the hidden state at decoding step t. At inference, only the encoder is used to produce the embedding for the input text. We train NaSE as the TSDAE encoder, with the following adjustment. We initialize NaSE with the

[5] We do not use a fixed length, accounting for the varying expressivity of characters in different languages.

[6] We set $K = 15\%$ for WikiNews, and $K = 3\%$ for the other four sources.

[7] We use 256 permutations and an n-gram size of 5. Following manual inspection, we set the de-duplication similarity threshold to 0.9.

[8] Fig. G.1 shows the distribution of texts across language pairs.

pretrained weights of the popular, widely used multilingual SE LaBSE (Feng et al., 2022).[9]

Our second training objective is domain-specific sequence-to-sequence machine translation (MT), for which we leverage the parallel data from PolyNewsParallel. For a given translation pair, we treat the input in the source language to be the 'corruption' $\tilde{x}$ of the *target* sentence x in the target language, which is to be 'reconstructed'. Using these two objectives, DAE and MT, we train four different variants of NaSE (as the resulting encoder of the corresponding sequence-to-sequence model): **(1) NaSE$_{\textbf{DAE}}$** reconstructs the original input sentence from its corruption; **(2) NaSE$_{\textbf{MT}}$** generates the *target*-language translation of the *source*-language input sentence; **(3) NaSE$_{\textbf{DAE + MT}}$** combines the two objectives, by randomly choosing either reconstruction or translation for each batch; for parallel data, the reconstruction objective is applied independently on both source- and target-language sentences; **(4) NaSE$_{\textbf{DAE} \rightarrow \textbf{MT}}$** is trained sequentially, first on reconstruction, and then on translation, i.e., we continue training the **NaSE$_{\textbf{DAE}}$** encoder for translation on parallel data. If not specified otherwise, with just **NaSE** we refer to the NaSE$_{\text{DAE} \rightarrow \text{MT}}$ variant. We train **NaSE$_{\textbf{DAE}}$** on PolyNews, and **NaSE$_{\textbf{MT}}$** and **NaSE$_{\textbf{DAE + MT}}$** on PolyNewsParallel. **NaSE** (i.e., NaSE$_{\text{DAE} \rightarrow \text{MT}}$), as a sequential combination of both training procedures, is trained first on PolyNews, and then on PolyNewsParallel.

9.4.2 Training Details

Training Data. The distributions of both PolyNews and PolyNewsParallel are heavily skewed across languages, and the number of news texts for some low-resource languages is particularly very limited.[10] We thus follow common practice and smoothen the per-language distribution when sampling for model training (Arivazhagan et al., 2019; Conneau et al., 2020; Xue et al., 2021). We first sample only languages and language pairs that contain at least 100 texts in PolyNews and PolyNewsParallel, respectively. We then sample texts from language L by sampling from the modified distribution $p(L) \propto |L|^{\alpha}$, with $|L|$ as the number of examples in L, and α as the smoothing rate hyperparameter ($\alpha < 1$ upsamples low-resource, and downsamples high-resource languages). We set α to 0.3, the value that was found to balance the performance between high- and low-resource languages well (Conneau et al., 2020; Xue et al., 2021).

Training Settings. We follow Wang et al. (2021a) and use token deletion with a ratio of 0.6 as the discrete corruption for the DAE variants of NaSE. In all training setups, we tie the encoder and decoder parameters, and initialize them with LaBSE weights. We train each model variant for 50K steps with a learning rate of $3e^{-5}$, using AdamW (Loshchilov and Hutter, 2018) as the optimizer. We checkpoint the model every 5K steps.

Validation Setup. We validate NaSE during training on cross-lingual news recommendation. Concretely, for each model checkpoint, we encode the clicked and candidate news using the *frozen* encoder. We adopt the late fusion approach (cf. Chapter 4), and replace the parameterized user encoder with mean-pooling of dot-product scores between the embeddings of the candidate and the clicked news. Eliminating a parameterized user encoder has two benefits: (1) it increases the computational efficiency of NaSE validation, as we do not have to train the NNR (i.e., its user encoder), and (2) it isolates NaSE (i.e., the news encoder), as the only component that affects the recommendation performance, eliminating the user encoder as a confounding factor.

Validation Data. A robust multilingual SE should produce good embeddings in many languages. We thus validate the quality of the NaSE embeddings during training—for the purposes of model (i.e., optimal checkpoint) selection—on the *small* variants of the English MIND (Wu et al., 2020e), and its multilingual xMIND (see Chapter 8) counterpart. xMIND comprises news articles from MIND machine-translated into 14 languages. To measure recommendation performance, we combine

[9] Because we start from an SE rather than vanilla LM, we remove from TSDAE's architecture the layer that pools token-level representation into a sentence embedding.

[10] E.g., only 100 texts in PolyNews for Tigrinya or Kurdish.

the multilingual news articles from xMIND with the user click behavior data from MIND; our final validation set covers user behaviors from the last day of the MIND training set. Our resulting validation corpus covers 15 linguistically diverse languages, offering a more realistic estimate of the quality of multilingual sentence embeddings produced by NaSE for XLT in news recommendation.

9.5 Experimental Setup

Zero-shot cross-lingual news recommendation is the task for which we primarily develop NaSE, and thus, the downstream task on which we evaluate it. In all experiments, we assume only monolingual news consumption, i.e., that each user reads news only in one language and, accordingly, also receives recommendations in one language.

Neural News Recommenders. We evaluate four architecturally diverse NNRs which showed promising results on the xMIND dataset (cf. Chapter 8): (1) *NAML* (Wu et al., 2019a), (2) *MINS* (Wang et al., 2022c), (3) *CAUM* (Qi et al., 2022b), and (4) *MANNeR* (cf. Chapter 7).[11] Additionally, we consider three simpler yet competitive baselines: (5) *LFRec-CE*, (6) *LFRec-SCL*, and (7) *NAML$_{CAT}$*. NAML, MINS, and CAUM were designed to encode textual information using pretrained word embeddings (Pennington et al., 2014), contextualized with CNNs (Kim, 2014) or attention layers (Bahdanau et al., 2015), whereas LFRec-CE, LFRec-SCL, and MANNeR use a pretrained LM as their news encoder. In addition to the news text, NAML, MINS, and CAUM leverage category information, while MANNeR and CAUM encode named entities extracted from the title and abstract of the news. Lastly, NAML$_{\text{CAT}}$ is a text-agnostic variant of NAML that learns news embeddings purely as randomly initialized and fine-tuned category vectors. Following Wu et al. (2021b), we replace the original news encoders of NAML, MINS, and CAUM with a pretrained LM in order to enable multilingual recommendation, and to ensure a fair comparison between models.[12] Concretely, we experiment with (1) the pretrained multilingual LM XLM-RoBERTa$_{\text{Large}}$ (Conneau et al., 2020), (2) the BERT-based multilingual SE LaBSE (Feng et al., 2022), and (3) our proposed NaSE.

The recommenders also differ in the way they model users. Specifically, NAML (Wu et al., 2019a) encodes their preferences using additive attention (Bahdanau et al., 2015), while MINS (Wang et al., 2022c) combines multi-head self-attention (Vaswani et al., 2017) with a multi-channel GRU-based recurrent network (Cho et al., 2014) and additive attention. The more complex CAUM (Qi et al., 2022b) uses a candidate-aware self-attention network to learn long-term user preferences, and a candidate-aware CNN to capture the users' short-term interests. It obtains the final user embeddings by attending over the two intermediate representations. In contrast to these models, MANNeR (cf. Chapter 7), LFRec-CE and LFRec-SCL resort to the non-parameterized late fusion approach introduced in Chapter 4 for aggregating click behaviors. With the exception of MANNeR and LFRec-SCL, which minimize the supervised contrastive loss (SCL, Khosla et al., 2020), the remaining recommenders are trained by minimizing the standard cross-entropy (CE) loss.

Data. We conduct experiments on the *small* variants of the English MIND (Wu et al., 2020e) and the multilingual xMIND (cf. Chapter 8) datasets. As mentioned in § 9.4.2, we couple the news texts from xMIND with the click behavior data from MIND, via news IDs. For details on the corresponding MIND statistics, we refer the reader to § 3.2.2.

Fine-Tuning Details. In all experiments that require LM fine-tuning, we update only the LM's last four layers.[13] We set the maximum click history length to 50, and sample four negatives per positive in training, as per Wu et al. (2022b). We tune the main hyperparameters of all NNRs

[11] Only the base version with the `CR-Module`, without any `A-Module` for aspect-based diversification.

[12] NAML, MINS, and CAUM pool token embeddings with an attention layer to obtain the sentence embedding from the LM, whereas MANNeR uses the vector of the `CLS` token.

[13] In early experiments with XLM-RoBERTa$_{\text{Base}}$, fine-tuning the whole LM did not bring gains compared to updating only the last four layers. For computational efficiency, we thus keep the bottom eight layers of LaBSE and NaSE, and the bottom 20 layers of XLM-RoBERTa$_{\text{Large}}$, frozen.

using grid search. More specifically, we search for the optimal learning rate in $\{1e^{-3}, 1^{e-4}, 1^{e-5}\}$, finding $1e^{-5}$ to be the most suitable value for all NNRs. We optimize the number of heads in the multi-head self-attention networks of NAML, MINS, CAUM in $[8, 12, 16, 24, 32]$, and the query vector dimensionality in the additive attention network of NAML and MINS in $[50, 200]$, with a step of 50. We use the following best performing settings: 32 attention heads for NAML and MINS, 8 for CAUM when coupled with XLM-RoBERTa$_{\text{Large}}$, and 24, 16, and 8 attention heads for NAML, MINS, and CAUM, respectively, when using LaBSE or NaSE as the news encoder's backbone. Moreover, we use a query vector of dimension 200 for NAML and 50 for MINS in combination with XLM-RoBERTa$_{\text{Large}}$, whereas for SE-based news encoders we use a dimensionality of 50 for NAML and 100 for MINS. We find the optimal temperature of 0.38 for the supervised contrastive loss in MANNeR and LFRec-SCL by sweeping the interval $[0.1, 0.5]$, with a step of 0.02. We set all the remaining model-specific hyperparameters to the optimal values reported in the respective papers. We train the models for 10 epochs, with a batch size of 8 for the SE-based variants, and 4 for the XLM-RoBERTa$_{\text{Large}}$-equipped NNRs. We train using mixed precision and the Adam optimizer (Kingma and Ba, 2014). We repeat each experiment three times with the seeds $\{42, 43, 44\}$, set with PyTorch Lightning's `seed_everything`, and report the mean and standard deviations for common metrics: AUC, MRR, and NDCG@10.[14] We train NaSE using Sentence Transformers (Reimers and Gurevych, 2019) and PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) on a cluster with virtual machines, on single NVIDIA A100 40GB GPUs.[15] We conduct all experiments with the NewsRecLib library (cf. § 3.3.2), on the same cluster, on single NVIDIA A40 48GB GPUs.

9.6 Results and Discussion

We compare the NNRs' performance for news encoders with different LM/SE backbones in cross-lingual transfer for news recommendation. In this setup, the user history and candidates during training are *monolingual* and in *English* only. At inference, both the user history and the candidate news are solely in one of the 14 *target languages* of xMIND.

Fine-Tuning News Encoders. We first investigate the standard recommendation, in which task-specific data (i.e., click behavior information and impressions) are used to train the NNR. Table 9.1 displays the performance of the recommenders when the underlying news encoders are fine-tuned (i.e., the weights of the backbone LM or SE are updated) for news recommendation, on English data.

While SE-powered NNRs outperform the text-agnostic baseline NAML$_{\text{CAT}}$, the NNRs relying on XLM-RoBERTa$_{\text{Large}}$ sometimes underperform this simple baseline. Similarly, XLM-RoBERTa$_{\text{Large}}$ underperforms LaBSE and NaSE on both English recommendation and in ZS-XLT, regardless of the NNR in which it is used. We believe this is because the XLM-RoBERTa$_{\text{Large}}$-based encoder first needs to learn how to aggregate token representations into sentence embeddings. These results clearly render (multilingual) SEs, pretrained to produce robust sentence embeddings, beneficial for news recommendation. SE-based NNRs achieve similar performance with LaBSE and NaSE, both in English, and in ZS-XLT on xMIND, i.e., NaSE does not bring gains over LaBSE from which we derived it. We argue that this is due to the fact that fine-tuning on news recommendation also leads to domain adaptation: LaBSE itself becomes sufficiently specialized for the news domain through large-scale recommendation fine-tuning, compensating for NaSE's task-agnostic domain adaptation.

Lastly, we note that the simple baselines LFRec-CE, and in particular LFRec-SCL, exhibit strong recommendation performance, often surpassing more complex models like CAUM or MINS, which have richer input (e.g., topical categories, named entities) and parameterized user encoders. This is in line with the findings of Chapter 4 and questions the need for complex parameterized user encoders.

[14] For brevity, we omit results for NDCG@5, as they exhibit the same patterns as NDCG@10.

[15] Code available at https://github.com/andreeaiana/nase. The implementation also provides the steps to create the PolyNews and PolyNewsParallel datasets.

Model	# Parameters (M)		AUC			MRR			NDCG@10		
	Trainable	Total	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ
NAML_{CAT}	0.387	0.387	$55.46_{\pm 0.18}$		0.00	$31.12_{\pm 0.56}$		0.00	$35.81_{\pm 0.59}$		0.00
$\text{LFRec-CE}_{\text{XLM-RoBERTa}_{\text{Large}}}$	307	559	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	0.00	$35.71_{\pm 0.70}$	$33.05_{\pm 0.60}$	-7.45	$40.32_{\pm 0.77}$	$37.66_{\pm 0.65}$	-6.59
$\text{LFRec-CE}_{\text{LaBSE}}$	414	470	$62.89_{\pm 1.50}$	$62.20_{\pm 1.08}$	-1.10	$36.71_{\pm 0.53}$	$35.65_{\pm 0.50}$	-2.90	$41.30_{\pm 0.60}$	$40.15_{\pm 0.56}$	-2.77
$\text{LFRec-CE}_{\text{NaSE}}$	414	470	$65.19_{\pm 0.23}$	$63.00_{\pm 1.71}$	-3.35	$36.24_{\pm 0.47}$	$33.39_{\pm 1.07}$	-7.85	$40.74_{\pm 0.44}$	$37.92_{\pm 1.13}$	-6.93
$\text{LFRec-SCL}_{\text{XLM-RoBERTa}_{\text{Large}}}$	307	559	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	0.00	$29.12_{\pm 0.77}$	$27.88_{\pm 0.84}$	-4.27	$33.28_{\pm 0.81}$	$31.96_{\pm 0.90}$	-3.96
$\text{LFRec-SCL}_{\text{LaBSE}}$	414	470	$68.94_{\pm 0.87}$	$67.29_{\pm 0.75}$	-2.39	$36.45_{\pm 0.65}$	$35.04_{\pm 0.53}$	-3.87	$41.14_{\pm 0.60}$	$39.65_{\pm 0.47}$	-3.62
$\text{LFRec-SCL}_{\text{NaSE}}$	414	470	$68.63_{\pm 1.23}$	$66.73_{\pm 0.91}$	-2.76	$36.59_{\pm 0.87}$	$34.78_{\pm 0.61}$	-4.95	$41.07_{\pm 0.82}$	$39.35_{\pm 0.52}$	-4.17
$\text{NAML}_{\text{XLM-RoBERTa}_{\text{Large}}}$	312	564	$55.05_{\pm 4.35}$	$53.54_{\pm 3.28}$	-2.72	$34.36_{\pm 1.28}$	$32.60_{\pm 0.69}$	-5.12	$38.90_{\pm 1.40}$	$37.04_{\pm 0.90}$	-4.77
$\text{NAML}_{\text{LaBSE}}$	414	471	$51.59_{\pm 0.71}$	$51.93_{\pm 0.65}$	+0.66	$36.63_{\pm 0.68}$	$35.91_{\pm 0.51}$	-1.98	$41.25_{\pm 0.71}$	$40.42_{\pm 0.49}$	-1.99
$\text{NAML}_{\text{NaSE}}$	414	471	$53.65_{\pm 0.90}$	$51.06_{\pm 0.54}$	-4.83	$36.97_{\pm 0.68}$	$34.38_{\pm 0.95}$	-7.00	$41.62_{\pm 0.66}$	$38.94_{\pm 0.93}$	-6.43
$\text{MINS}_{\text{XLM-RoBERTa}_{\text{Large}}}$	316	568	$52.18_{\pm 2.86}$	$53.43_{\pm 3.00}$	+2.40	$31.64_{\pm 1.98}$	$30.95_{\pm 2.19}$	-2.19	$36.12_{\pm 2.05}$	$35.39_{\pm 2.24}$	-4.77
$\text{MINS}_{\text{LaBSE}}$	416	473	$63.70_{\pm 1.02}$	$62.31_{\pm 0.77}$	-2.19	$34.31_{\pm 0.80}$	$33.58_{\pm 0.58}$	-2.11	$39.06_{\pm 0.69}$	$38.22_{\pm 0.46}$	-2.17
$\text{MINS}_{\text{NaSE}}$	416	473	$62.33_{\pm 0.95}$	$60.81_{\pm 1.23}$	-2.44	$34.54_{\pm 0.89}$	$33.53_{\pm 0.67}$	-2.92	$39.36_{\pm 0.86}$	$38.35_{\pm 0.62}$	-2.58
$\text{CAUM}_{\text{XLM-RoBERTa}_{\text{Large}}}$	316	568	$58.67_{\pm 0.41}$	$59.11_{\pm 0.74}$	+0.74	$30.01_{\pm 2.30}$	$30.04_{\pm 1.98}$	+0.08	$34.57_{\pm 2.29}$	$34.57_{\pm 2.07}$	0.00
CAUM	417	474	$64.92_{\pm 0.83}$	$63.52_{\pm 0.80}$	-2.16	$34.36_{\pm 0.40}$	$33.53_{\pm 0.69}$	-2.69	$39.40_{\pm 0.21}$	$38.34_{\pm 0.54}$	-2.70
$\text{CAUM}_{\text{NaSE}}$	417	474	$64.54_{\pm 0.53}$	$62.54_{\pm 0.37}$	-3.10	$35.24_{\pm 0.40}$	$33.97_{\pm 0.27}$	-3.61	$40.09_{\pm 0.35}$	$38.59_{\pm 0.29}$	-3.74
$\text{MANNeR}_{\text{XLM-RoBERTa}_{\text{Large}}}$	310	562	$60.76_{\pm 9.32}$	$59.96_{\pm 8.62}$	-1.31	$35.54_{\pm 0.59}$	$34.02_{\pm 0.65}$	-4.27	$39.84_{\pm 0.47}$	$38.41_{\pm 0.76}$	-3.59
$\text{MANNeR}_{\text{LaBSE}}$	415	472	$68.31_{\pm 1.26}$	$66.89_{\pm 1.21}$	-2.09	$36.52_{\pm 0.65}$	$35.24_{\pm 0.52}$	-3.50	$41.11_{\pm 0.65}$	$39.81_{\pm 0.48}$	-3.16
$\text{MANNeR}_{\text{NaSE}}$	415	472	$67.22_{\pm 0.40}$	$65.29_{\pm 0.68}$	-2.87	$35.41_{\pm 0.61}$	$33.86_{\pm 0.44}$	-4.36	$39.88_{\pm 0.41}$	$38.44_{\pm 0.36}$	-3.59

Table 9.1: ZS-XLT recommendation performance with a fine-tuned news encoder (LM/SE). For each model, we report its size in terms of millions of trainable and total parameters, and performance (i) on the English MIND (ENG), (ii) averaged across all 14 target languages in xMIND (AVG), and (iii) the relative percentage difference between average ZS-XLT and ENG performance (%Δ). Reported performance is averaged over three runs. Subscripts denote standard deviation.

Frozen News Encoders. Fine-tuning news encoder backbones (LMs with hundreds of millions of parameters) on large-scale recommendation data can be prohibitively expensive for many practitioners without access to large computational resources (e.g., GPUs). We thus next analyze how NNRs perform with frozen news encoders (i.e., no updates to LM/SE), allowing updates only to other (fewer) trainable parameters of the news recommenders. Specifically, the model can now only learn how to encode other input features (e.g., categories), or how to aggregate the news embeddings into a user-level encoding, if equipped with parameterized user encoders. For most models, freezing the news encoder reduces the number of trainable parameters by two orders of magnitude.

Table 9.2 summarizes the recommendation performance of the NNRs with frozen news encoders. Unsurprisingly, $\text{XLM-RoBERTa}_{\text{Large}}$-based recommenders yield the weakest performance across all languages, as the LM itself cannot be tuned to encode token sequences. Out-of-the-box NaSE embeddings substantially outperform LaBSE in most cases cases: e.g., for NDCG@10 by 2.58% on English, and 4.17% cross-lingually (averaged over all 14 xMIND languages), averaged across all NNRs. The performance gap between LaBSE and NaSE becomes smaller when the NNR uses a more complex *trainable* user encoder (e.g., MINS): the user encoder parameters take over the domain-specialization task in large-scale fine-tuning.

Two key aspects point to successful (task-agnostic) domain-specialization of NaSE. **1)** We observe that for models without trainable user encoders the relative performance loss in ZS-XLT compared to English performance is less pronounced with NaSE as news encoder than with LaBSE. As target-language data is often scarce in many real-world applications, closing the gap to performance on the source language on which the recommender is trained is crucial for multilingual news recommendation. **2)** Unlike $\text{XLM-RoBERTa}_{\text{Large}}$ or LaBSE, who both benefit from large-scale recommendation fine-tuning, NaSE's performance when frozen is much closer to its fine-tuned performance. This calls into question the current paradigm of performing expensive supervised fine-tuning of the NNR's news encoder.

Most importantly, coupled with a frozen NaSE NE, LFRec-SCL – a model with no other trainable parameters – achieves state-of-the-art performance over more complex, trainable NNRs. On the one hand, this demonstrates the effectiveness of the news-specialized NaSE encoder (trained in a task-agnostic manner) over a general-purpose SE like LaBSE in news recommendation. On the

Model	# Parameters (M)		AUC			MRR			NDCG@10		
	Trainable	Total	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ
LFRec-SCL $_{\text{XLM-RoBERTa}_{\text{Large}}}$	0	559	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	-0.00	$24.82_{\pm 0.00}$	$24.64_{\pm 0.00}$	-0.72	$28.44_{\pm 0.00}$	$28.71_{\pm 0.00}$	+0.92
LFRec-SCL $_{\text{LaBSE}}$	0	471	$50.47_{\pm 0.00}$	$50.98_{\pm 0.00}$	+1.00	$30.96_{\pm 0.00}$	$28.99_{\pm 0.00}$	-6.35	$34.83_{\pm 0.00}$	$32.95_{\pm 0.00}$	-5.41
LFRec-SCL $_{\text{NaSE}}$	0	471	$64.19_{\pm 0.00}$	$63.79_{\pm 0.00}$	-0.62	$34.74_{\pm 0.00}$	$33.96_{\pm 0.00}$	-2.26	$39.01_{\pm 0.00}$	$38.23_{\pm 0.00}$	-2.00
NAML$_{\text{XLM-RoBERTa}_{\text{Large}}}$	4.9	564	$50.01_{\pm 0.01}$	$50.01_{\pm 0.01}$	0.00	$26.75_{\pm 1.11}$	$26.59_{\pm 0.92}$	-0.62	$31.06_{\pm 1.26}$	$30.93_{\pm 0.89}$	-0.42
NAML$_{\text{LaBSE}}$	0.156	471	$61.17_{\pm 0.01}$	$58.78_{\pm 0.70}$	-3.91	$34.57_{\pm 0.09}$	$33.86_{\pm 0.32}$	-2.05	$39.21_{\pm 0.05}$	$38.19_{\pm 0.35}$	-2.60
NAML$_{\text{NaSE}}$	0.156	471	$63.30_{\pm 0.14}$	$61.05_{\pm 0.33}$	-3.55	$33.89_{\pm 0.16}$	$32.73_{\pm 0.21}$	-3.43	$38.47_{\pm 0.10}$	$37.47_{\pm 0.18}$	-2.60
MINS$_{\text{XLM-RoBERTa}_{\text{Large}}}$	8.8	568	$49.93_{\pm 0.28}$	$50.30_{\pm 0.22}$	+0.75	$29.77_{\pm 0.42}$	$28.13_{\pm 1.49}$	-5.51	$33.96_{\pm 0.52}$	$32.74_{\pm 1.47}$	-3.57
MINS$_{\text{LaBSE}}$	2.7	473	$61.25_{\pm 1.19}$	$60.37_{\pm 1.38}$	-1.43	$32.69_{\pm 0.35}$	$32.00_{\pm 0.51}$	-2.09	$37.37_{\pm 0.39}$	$36.50_{\pm 0.55}$	-2.34
MINS$_{\text{NaSE}}$	2.7	473	$60.19_{\pm 0.64}$	$58.62_{\pm 0.81}$	-2.61	$32.23_{\pm 0.54}$	$31.65_{\pm 0.68}$	-1.80	$37.08_{\pm 0.55}$	$36.44_{\pm 0.70}$	-1.73
CAUM$_{\text{XLM-RoBERTa}_{\text{Large}}}$	8.2	559	$55.14_{\pm 0.61}$	$55.00_{\pm 0.67}$	-0.27	$28.55_{\pm 0.71}$	$28.40_{\pm 0.76}$	-0.53	$33.26_{\pm 0.59}$	$33.14_{\pm 0.69}$	-0.36
CAUM$_{\text{LaBSE}}$	3.7	474	$62.82_{\pm 0.77}$	$61.35_{\pm 0.74}$	-2.34	$33.79_{\pm 0.59}$	$33.00_{\pm 0.48}$	-2.33	$38.40_{\pm 0.56}$	$37.61_{\pm 0.46}$	-2.05
CAUM$_{\text{NaSE}}$	3.7	474	$64.40_{\pm 0.61}$	$62.69_{\pm 0.64}$	-2.65	$34.42_{\pm 0.44}$	$33.49_{\pm 0.37}$	-2.72	$39.15_{\pm 0.43}$	$37.98_{\pm 0.34}$	-2.98
MANNeR$_{\text{XLM-RoBERTa}_{\text{Large}}}$	2.1	562	$62.11_{\pm 1.17}$	$50.80_{\pm 0.30}$	-18.21	$30.88_{\pm 1.42}$	$23.86_{\pm 0.27}$	-22.73	$35.56_{\pm 1.18}$	$27.78_{\pm 0.25}$	-21.88
MANNeR$_{\text{LaBSE}}$	1.7	472	$64.54_{\pm 0.97}$	$61.97_{\pm 0.85}$	-3.98	$33.96_{\pm 0.87}$	$31.45_{\pm 0.76}$	-7.37	$38.51_{\pm 0.77}$	$35.82_{\pm 0.67}$	-6.98
MANNeR$_{\text{NaSE}}$	1.7	472	$65.51_{\pm 1.07}$	$64.14_{\pm 0.78}$	-2.08	$34.68_{\pm 0.89}$	$33.45_{\pm 0.72}$	-3.54	$39.13_{\pm 0.77}$	$37.94_{\pm 0.62}$	-3.05

Table 9.2: ZS-XLT recommendation performance with a frozen news encoder (LM/SE). For each model, we report its size in terms of millions of trainable and total parameters, and performance (i) on the English MIND (ENG), (ii) averaged across all 14 target languages in xMIND (AVG), and (iii) the relative percentage difference between average ZS-XLT and ENG performance (%Δ). Reported performance is averaged over three runs. Subscripts denote standard deviation.

other hand, it proves that LFRec-SCL can produce good recommendations in true cold-start setup, where no news or user data is available (for training a parameterized NNR). This is in contrast to all the other models that require task-specific data to learn meaningful user representations, which are then to compute the final recommendation scores. Notably, this means that with LFRec-SCL with (frozen) NaSE we obtain a state-of-the-art news recommendation performance without the need for any task-specific training for news recommendation. LFRec-SCL with (frozen) NaSE outperforms more complex models (NAML, MINS, CAUM), with user encoders trained for recommendation, in terms of ranking (i.e., MRR, NDCG@k), and performs on-par or better in terms of classification (i.e., AUC). These results suggest that domain-specialization of a multilingual SE (i.e., NaSE) removes the need for supervised training of NNRs.

Few-Shot Recommendation. To corroborate the previous findings, we also investigate the performance of the NNRs when the underlying news encoder is fine-tuned on just a few task-specific examples, namely in *few-shot recommendation*. Here, we assume only a handful of task-specific examples (i.e., impressions) for training the NNRs. In terms of XLT, we stick to zero-shot XLT for recommendation, i.e., we assume that the few training instances that we have are all in the source language, i.e., English.[16] Fig. 9.2 shows the results for 10, 50, and 100 shots, i.e., the number of impressions used to train the NNR. We again observe that, when fine-tuned on small amounts of data, NaSE as the news encoder generally outperforms LaBSE, especially for fewer shots (10 and 50). In 10-shot recommendation, NaSE's relative gain in terms of NDCG@10 over LaBSE ranges from 3.96% (averaged over 14 target languages) and 0.21% on English (for NAML) to 12.54% and 7.38% (for LFRec-SCL), respectively. The differences between the two decrease with more training data (100 shots). Besides cold starts, these results render NaSE effective in ZS-XLT recommendation in realistic low-data setups.

Impact of Training Strategy. Lastly, we ablate NaSE's performance for different training strategies. Fig. 9.3 shows the ranking performance (i.e., NDCG@10) of LFRec-SCL with frozen NaSE embeddings. We find that the denoising auto-encoding pretraining objective (DAE) underperforms all other NaSE configurations. Nonetheless, even LFRec-SCL with frozen NaSE$_{\text{DAE}}$ embeddings achieves between 0.19% (on THA) and 12.90% (on FIN) relative improvements over the model using frozen LaBSE embeddings, outperforming it on 12 out of 15 languages. The translation objective seems

[16] This should not be confused with few-shot XLT for recommendation, which assumes (few) target-language instances in the user histories during training, but overall relies on large amounts of task-specific training data (see Chapter 8).

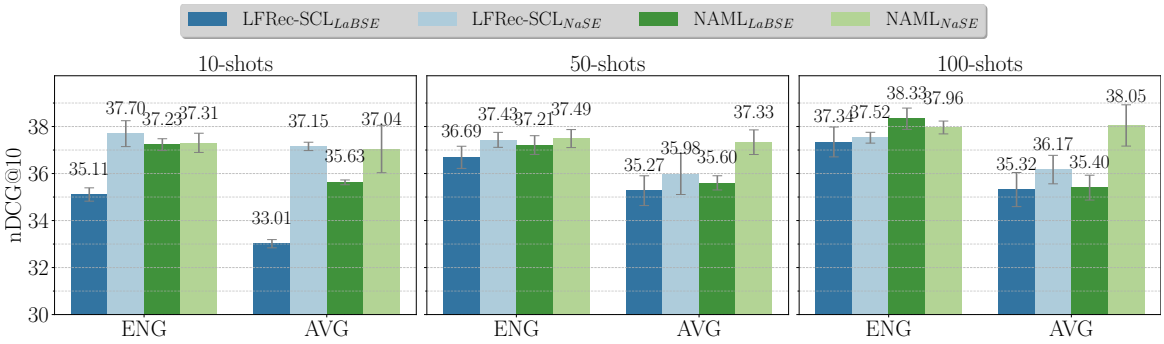


Figure 9.2: ZS-XLT performance in few-shot recommendation (i.e., for different number of impressions used in training). For each model, we report average performance across three runs (i) on the English xMIND dataset (denoted ENG), and (ii) averaged across all 14 target languages of xMIND (denoted AVG).

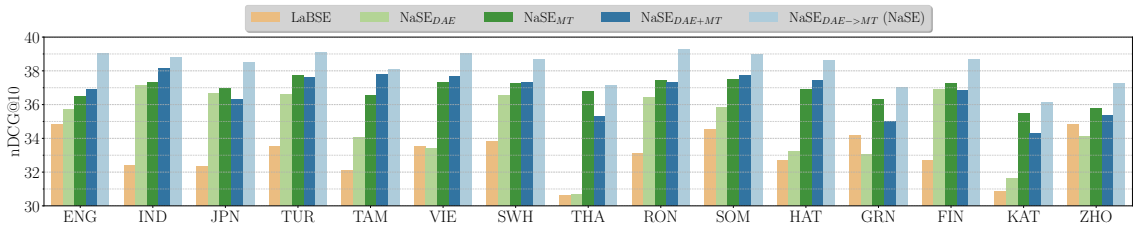


Figure 9.3: Ranking performance (NDCG@10) of LFRec-SCL with frozen NaSE embeddings for different training strategies of NaSE versus LFRec-SCL with frozen LaBSE embeddings, over English (ENG) and the 14 languages in xMIND.

highly effective for all languages: NaSE$_{\text{MT}}$ brings up to 19.99% and 20.22% improvements relative to NaSE$_{\text{DAE}}$, and LaBSE, respectively. Interestingly, the gains seem largest for languages not present in PolyNewsParallel on which we train NaSE (e.g., THA) and those not seen by LaBSE in pretraining (e.g., GRN): this hints at positive cross-lingual transfer of domain-specialization via the MT training.

Training on both DAE and MT slightly improves DAE-only specialization for most languages. However, *sequentially* combining the two (NaSE$_{\text{DAE}\rightarrow\text{MT}}$; referred to with just NaSE) brings further gains and overall best performance. We believe that this is due to the sequential training being exposed to more data, i.e., both PolyNews (in the DAE stage) and PolyNewsParallel (in the MT stage). Compared to LaBSE, NaSE in this setup yields relative improvements ranging from 7.05% on ZHO, to 11.99% on ENG, and 21.33% on THA.

It is worth emphasizing that news-specialization of NaSE with a translation-based objective benefits also languages *not* included in PolyNewsParallel (e.g., THA, HAT, KAT), and even languages not present in LaBSE's large-scale parallel pretraining corpus (e.g., GRN). This points to the language-agnostic nature of news-specific domain specialization. This finding is in line with results from prior work in machine translation (Tran et al., 2020; Duquenne et al., 2023), which showed that modest amounts of parallel data (as in our PolyNewsParallel) improve the cross-lingual semantic alignment in multilingually trained encoder-decoder models. This then results in performance gains even for languages not present in the parallel corpora.

9.7 Conclusion

Current neural news recommenders based on multilingual language models (i) suffer from substantial performance losses in ZS-XLT recommendation, and (ii) usually require expensive fine-tuning of

the language model used as the news encoder backbone: such fine-tuning is often infeasible, e.g., in cold-start setups without click-behavior data. In this Chapter, we proposed NaSE, a news-adapted sentence encoder obtained through domain specialization of a pretrained multilingual sentence encoder. To this end, we compiled and leveraged two multilingual news-specific corpora, PolyNews and PolyNewsParallel. Our findings question the effectiveness of supervised fine-tuning for news recommendation. As an efficient solution, we proposed a simple and strong baseline based on frozen NaSE embeddings and late click behavior fusion that achieves state-of-the-art performance in ZS-XLT in true cold-start and few-shot news recommendation. In the next and final Chapter, we summarize and conclude the findings and contributions presented in this thesis.

Part V

CONCLUSIONS AND PERSPECTIVES

Chapter 10

Conclusions and Perspectives

All men have stars, but they are not the same things for different people.
For some, who are travelers, the stars are guides.
For others they are no more than little lights in the sky.
For others, who are scholars, they are problems...

Antoine de Saint-Exupéry
«The Little Prince »

In this thesis, we proposed methods and resources to advance the development of responsible and inclusive personalized neural news recommender systems, as motivated in Section 1.1. In this Chapter, we first summarize our key insights and findings (Section 10.1), and then conclude with a discussion of promising avenues for future research (Section 10.2).

10.1 Summary of Findings

Personalized news recommender systems play a fundamental role in today's information ecosystem, shaping how individuals access news and, ultimately, how they form opinions and worldviews. In recent years, research in this area has increasingly converged on neural approaches, driven by their ability to model rich textual content and complex user behavior. While core technical challenges in news recommendation persist (cf. Section 1.1), this methodological shift unfolded in parallel with growing attention to the broader societal implications of large-scale news recommender systems. Issues such as biased exposure, information diversity, and unequal access across languages have become central concerns, underscoring that effectiveness alone is insufficient. Developing news recommender systems that meet real-world deployment demands therefore requires addressing technical performance alongside questions of responsibility, inclusivity, and societal impact. In this thesis, we have argued for and adopted a *multifaceted approach* to the development of personalized recommendation technology in the news domain: one that seeks to advance efficiency while explicitly accounting for the societal effects of algorithmic news curation and the inequities introduced by linguistic diversity. Against the backdrop of this perspective, we identified and addressed four fundamental and complementary challenges: understanding the role of common architectural components, analyzing the effects of algorithmic bias, enabling flexible multi-aspect customization, and supporting multilingual news recommendation. In the following, we synthesize key findings and contributions arising from each of these challenges.

Analysis of Common Components (C1). The performance of neural news recommenders depends on the interaction of their core components, yet the effects of common design choices have remained poorly understood; we addressed this gap through two complementary analyses. First, in Chapter 4, we introduced a unified framework to systematically compare NNRs across candidate-aware user

modeling, click behavior fusion, and training objectives. Within this framework, we proposed a simple *late fusion* strategy for user modeling, showing that it substantially reduces model complexity while matching or surpassing the performance of more complex user encoders. We further demonstrated that supervised contrastive learning is an effective alternative to standard classification objectives, yielding competitive improvements in recommendation performance. Second, we proposed a multifaceted evaluation protocol that analyzes news and user encoder behavior beyond accuracy in terms of representational similarity and recommendation overlap (see Chapter 5). These results showed that architectural complexity is not uniformly required across all components of a NNR. Instead, performance is primarily driven by the semantic quality of news representations, while user encoders can be significantly simplified without sacrificing accuracy. Taken together, these findings provide a principled understanding of how common components contribute to neural news recommendation performance. They suggest that architectural complexity should be applied selectively – favoring more expressive news encoders where semantic modeling is critical, while keeping user encoders lightweight – and yield concrete guidelines for balancing model simplicity, efficiency, and effectiveness when designing NNRs for deployment.

Effects of Algorithmic Bias (C2). While algorithmic news curation alleviates information overload, it may also inadvertently reinforce selective exposure and propagate existing societal biases, particularly when optimized primarily for relevance. To examine how such effects manifest in neural news recommender systems, we investigated content-specific biases and their interaction with user behavior in politically sensitive contexts (Chapter 6). We conducted an aspect-based analysis of recommendation bias along multiple content dimensions, including affective, topical, and political characteristics. To enable this analysis, we introduced two curated datasets on migration-related news: the German GeNeG corpus and the bilingual German–English NeMig corpus, both enriched with fine-grained content annotations and real-world user interactions. Our results revealed a consistent tendency across models to recommend negatively framed content. More critically, for text-based recommenders, these biases correlated with users' pre-existing preferences, indicating a risk of opinion amplification and reduced exposure to diverse perspectives, particularly when recommenders are optimized solely for relevance. We further showed that personalization effects are more strongly aligned with topical and political dimensions than with affective characteristics such as sentiment. Overall, these findings provide empirical evidence that NNRs can reinforce selective exposure and content-level biases, underscoring the need to evaluate and design NNRs with explicit consideration of their societal effects beyond accuracy-centric metrics.

Multi-Aspect Customization of Recommendations (C3). While prior work has proposed diversity- and fairness-aware recommendation models to mitigate the negative effects of algorithmic news curation, such approaches typically target a single aspect or objective at a time and require retraining whenever either the overall recommendation goal or the specific diversification subgoals within it change. This rigidity limits their applicability in real-world settings, where users differ in how they trade off relevance against multiple, potentially competing criteria. To address this crucial limitation, in Chapter 7 we proposed MANNeR, a modular framework for multi-aspect neural news recommendation that supports efficient, user-controllable customization beyond pure relevance. Our model employs metric-based learning to train independent news encoders for different aspects (e.g., content, topic, sentiment) and combines their aspect-specific similarity scores linearly at inference time. This modular design enables the definition of ad-hoc ranking functions per user, allowing dynamic adjustment of trade-offs between content relevance, aspect-based diversification, and aspect-based personalization without model retraining. Importantly, this flexibility provides a practical mechanism to address the content-level biases identified in the previous challenge (cf. **C2** and Chapter 6) by enabling explicit control over competing objectives. We showed that MANNeR consistently outperforms state-of-the-art NNRs across standard content-based recommendation, as well as on single- and multi-aspect diversification and personalization tasks. Moreover, MANNeR's modular structure enables seamless extension to new aspects and robust cross-lingual transfer of the

aspect-specific news encoders when instantiated with a multilingual language model. These results demonstrate that flexible, multi-aspect, user-adaptive recommendation can seamlessly be achieved within a single unified framework. At the same time, MANNeR's modularity is particularly well suited to multilingual contexts, where data availability and user preferences vary across languages and where retraining separate models per language might often be infeasible.

Multilinguality (C4). News content is produced and consumed in virtually all societies, yet language fundamentally governs which content users can access and how they engage with it. Despite growing linguistic diversity online and increasing multilingual news consumption, research on neural news recommendation remains dominated by monolingual benchmarks and systems tailored to resource-rich languages, implicitly assuming monolingual user behavior. This mismatch reinforces the digital language divide (cf. 1.1.2), systematically disadvantaging speakers of low-resource languages and limiting the extent to which multilingual readers' broader linguistic capabilities can be supported by current NNRs. In the last part of this thesis, we advanced more inclusive neural news recommendation by explicitly addressing multilinguality in both evaluation resources and modeling approaches. In Chapter 8, we introduced xMIND, the first open, multilingual, multi-parallel benchmark for cross-lingual news recommendation, covering 14 geographically and linguistically diverse languages. Its multi-parallel structure enables controlled comparison of model performance across high- and low-resource languages and supports the simulation of multilingual news consumption. Using xMIND, we showed that existing NNRs, even when based on multilingual language models, exhibit severe performance degradation in zero-shot cross-lingual transfer, while few-shot adaptation yields only limited gains, particularly for bilingual users and linguistically distant languages. Building on these findings, in Chapter 9 we proposed NaSE, a news-adapted multilingual sentence encoder trained via domain-specific adaptation on two newly compiled multilingual news corpora. We demonstrated that NaSE consistently improves cross-lingual performance over general-purpose language models, especially in zero-shot and cold-start settings when its parameters are kept frozen, underscoring the importance of robust representations over task-specific fine-tuning in data-scarce scenarios. Based on this insight, we further established a simple yet strong baseline combining frozen NaSE representations with late fusion, achieving state-of-the-art results in cold-start and few-shot news recommendation. Overall, these contributions establish a principled foundation for multilingual news recommendation, revealing both the limitations of current approaches and showcasing effective paths forward. By pairing rigorous multilingual evaluation with domain-adapted representations, we showed that more equitable and scalable personalization across languages can be substantially improved even in data-scarce and cold-start settings.

To summarize, this thesis has argued for and demonstrated the importance of adopting a multi-faceted perspective in the design of efficient, inclusive, and responsible neural news recommender systems. To this end, we identified and addressed four complementary challenges, and introduced new evaluation resources and corpora, novel modeling frameworks, and conducted empirical analyses that together contribute to a more principled understanding of neural news recommendation while explicitly engaging with its societal implications.

At the same time, our work only scratches the surface of the complex technical, societal, and epistemic questions surrounding algorithmic news curation. Many of the proposed methods are grounded in textual news and interaction data from online news platforms, abstracting away from emerging forms of news consumption that increasingly shape public discourse, such as podcasts, multimodal and short-form video content, and highly dynamic creator-driven ecosystems. Moreover, while this thesis highlights the risks of algorithmic bias and selective exposure, it does not resolve the broader challenges posed by the growing prevalence of AI-generated content in news production, nor the normative questions surrounding responsibility, transparency, and power asymmetries in large-scale recommender systems. These limitations point to the need for a more holistic research agenda that spans data collection, model development, evaluation protocols, and governance considerations. In a rapidly evolving news ecosystem, marked by multimodal consumption patterns, generative

models as content producers, and increased societal scrutiny of algorithmic influence, developing truly inclusive and responsible news recommendation technology requires sustained efforts across all building blocks of recommendation pipelines. In the following section, we therefore outline perspectives for future research, charting promising avenues along the core axes of available resources, modeling paradigms, and broader societal and ethical implications.

10.2 Perspectives on Effective, Responsible, and Inclusive (News) Recommendation

Text Is *Not* All You Need. Arguably, most research on neural news recommendation has focused on textual content and text-derived features, largely overlooking other rich modalities. Yet, news consumption is shifting rapidly, with audiences increasingly engaging with video-based news, filmed podcasts, and short-form formats on platforms such as YouTube and TikTok, especially among younger users (Newman et al., 2025). Few works have incorporated visual content in news recommendation. Multimodal recommenders that leverage images alongside text have shown improvements in news representation and click prediction (Wu et al., 2021c; Xun et al., 2021; Han et al., 2022). Despite the growing importance of video and audio modalities, research remains scarce, and there is a notable lack of suitable benchmarks and modeling approaches, though early work on news video recommendation exists (Luo et al., 2008). Recent advances in multimodal recommendation in other domains provide methods readily adaptable to news. Techniques for fusing text, images, audio, and other modalities can improve recommendation quality, relevance and robustness (Lubos et al., 2023; Liu et al., 2024a; Xu et al., 2025). For example, multimodal benchmarks like ViLLA-MMBench (Nazary et al., 2025) illustrate the benefits of aligned audio-visual-text representations in media recommendation, while multimodal LLMs for vision and language have been successfully applied to video recommendation tasks (De Nadai et al., 2025). Embedding-based architectures such as SiBraR (Ganhör et al., 2024) enable robust fusion across heterogeneous modalities, supporting cold-start and missing-modality scenarios.

These developments, together with the evolving trends in news consumption, suggest that "text is not all you need". At the same time, incorporating additional modalities beyond text also raises new challenges for balanced and responsible recommendation, as visual and audiovisual content, particularly emotionally charged videos or imagery, can potentially amplify affective cues and influence user attention and engagement (Malitesta et al., 2023; Shang et al., 2024; Janghorbani and De Melo, 2023; Chen et al., 2025b). This observation begs the question of how NNRs should be redesigned to reflect the multimodal reality of contemporary news consumption. Neural news recommendation should embrace a broader multimodal perspective, leveraging methods from multimodal NLP, IR, and computer vision – including multimodal encoders, cross-modal representation learning, and deep fusion architectures – to integrate video, audio, and image signals alongside text. Advancing multimodal news recommenders will require both new benchmarks and modeling approaches to capture interactions across heterogeneous modalities, enabling more expressive, context-aware, and user-aligned recommendations that reflect how users increasingly consume news beyond text, while accounting for modality-specific biases and affective signals. Integrating multiple modalities may also help mitigate biases toward text-heavy content (cf. § 6.4.2) and promote more equitable exposure across diverse news formats.

Towards Richer and More Representative Data Resources. Data availability and quality are foundational to the development of effective, inclusive, and responsible news recommender systems. To date, much of the neural news recommendation research has relied on a small number of static, text-centric benchmarks in a limited set of languages, most prominently Adressa in Norwegian Gulla et al. (2017) and MIND in English (Wu et al., 2020e). While these datasets have enabled substantial methodological progress, they offer only a narrow view of real-world news ecosystems,

often underrepresenting politically sensitive or hard news and constraining the study of bias, exposure, and normative objectives beyond accuracy. This limitation is particularly consequential in news recommendation, where rapid temporal drift and increased societal sensitivity distinguish the domain from other recommendation contexts. As a result, dataset representativeness, freshness, and coverage are especially critical for both empirical validity and responsible deployment. More recent efforts, such as NPR (Lucas et al., 2023), EB-NeRD (Kruse et al., 2024), and the datasets introduced in this thesis (i.e., GeNeG, NeMig, cf. Chapter 6), move toward richer content and annotation schemes that better support the study of framing, ideology, and exposure effects. Nevertheless, these resources remain largely monolingual, Western-centric, and predominantly text-based. Even multilingual benchmarks like xMIND extend linguistic coverage but still capture only a partial and culturally skewed snapshot of contemporary news consumption.

Looking forward, progress in responsible and inclusive news recommendation will require datasets that are more representative, richer in signals, and dynamically maintained. This entails (i) broader linguistic and geographic coverage to support multilingual and cross-lingual news consumption, (ii) richer item representation spanning multiple modalities, and (iii) user interaction signals that go beyond clicks to capture contextual and intent-aware factors, as explored in datasets such as NeMig (see Section 6.5) and IDEA (Heitz et al., 2024). At the same time, the fast-moving nature of news raises fundamental challenges for dataset longevity and evaluation validity. Emerging infrastructures such as POPROX (Burke and Ekstrand, 2025), which enable controlled online experimentation on live news streams, illustrate one promising direction, though they raise trade-offs with reproducibility and accessibility. A complementary path lies in hybrid benchmarking strategies that combine large-scale static training corpora (e.g., MIND, EB-NERD) with smaller, periodically refreshed evaluation sets from ongoing news streams. Such approaches, increasingly explored across recommendation (Higley et al., 2025), IR (Galuscáková et al., 2023; Chen et al., 2025a; Thakur et al., 2025), and NLP (Kiela et al., 2021; Ying et al., 2024), help balance more realistic evaluation scenarios with comparability while mitigating the risk of test-set contamination and benchmark saturation.

Crucially, advancing inclusive and responsible news recommendation necessitates closer collaboration between academia and industry, particularly around dataset creation and model evaluation. In the news domain, industry actors are often the only stakeholders with access to temporally up-to-date interaction logs, multilingual content streams, and real-world deployment capabilities that are essential for studying the performance of NNRs, exposure dynamics, and societal impact at scale. Academia, in turn, contributes controlled experimental design and frameworks for analyzing bias, diversity, and normative trade-offs that are difficult to address in purely production-driven settings. Bridging these complementary strengths through shared benchmarks, privacy-preserving data releases, or controlled evaluation infrastructures would enable more realistic and reproducible research, while supporting industry in developing news recommenders that are not only effective, but also accountable, transparent, and aligned with democratic values and societal responsibility.

AI-Generated Content and Emerging Societal Challenges. The increasing prevalence of AI in news highlights two distinct but interrelated roles of large language models: (i) as interfaces that curate and summarize human-authored content, and (ii) as active producers generating content, which may be partially or fully synthetic. Each role carries unique implications for news recommendation, including effects on user perception, misinformation, and fairness.

LLMs as Interfaces to Traditional News. The Reuters Digital News Report 2025 documents that AI chatbots and interfaces are emerging as a source of news for a non-trivial share of users, particularly younger audiences (see Section 1.1 and Fig. 1.2). These tools provide convenient access to curated news summaries and personalized highlights. However, research demonstrates that AI assistants frequently produce factual inaccuracies and sourcing errors in LLMs-produced news summaries, amplifying concerns about misinformation and erosion of public trust (Le Poidevin, 2025). In this role, LLMs primarily influence how users consume and interact with existing news, affecting engagement signals (e.g., clicks, dwell time) that feed into recommendation models.

Future work should investigate the user-centered implications of AI-mediated news interfaces. Key directions include (i) developing methods for trust calibration and transparency, such as provenance-aware labels[1] or confidence scores, to help users interpret LLM-curated summaries reliably; (ii) exploring interaction-aware recommendation, modeling how exposure to AI interfaces influences engagement signals, reading behavior, and long-term consumption patterns; (iii) assessing fairness and bias in user interaction with LLM-driven interfaces such as chatbots, to ensure that interface-mediated personalization does not unintentionally favor certain topics, sources, or demographic groups; (iv) designing explainable and conversational interfaces that provide feedback and reasoning for recommendations to improve user satisfaction and trust (Glassberg et al., 2025; Polleti et al., 2025); and (v) conducting longitudinal studies to examine long-term effects of LLM-mediated news interfaces on user perception, diversity of content exposure, and reliance on LLM-curated summaries. Collectively, these directions aim to enable responsible, user-aligned news recommendation in settings where AI increasingly mediates access to traditional news content.

LLMs as News Producers. Generative AI is increasingly used to shape news production, with a growing fraction of articles being automatically generated, sometimes without transparent disclosure (Simon, 2024; Russell et al., 2025). At the same time, users remain ambivalent about the reliability of AI-produced content, often perceiving it as less trustworthy than human journalism (Newman et al., 2025). Stylistic analyses reveal systematic differences in readability, formality, and writing uniformity between AI-generated and human-authored news (Ansari et al., 2025). These trends have direct implications for news recommendation: generative AI may alter the candidate item space, shift the distribution of content available for recommendation, and amplify feedback loops in which synthetic material disproportionately influences interaction histories and model training. Related work has begun to formalize this source bias in recommendation, showing how uncorrected integration of AI-generated content can contaminate training data and reinforce echo chambers (Zhou et al., 2025b).

Future work should therefore explicitly incorporate AI-generated content into news recommendation pipelines, both in benchmarking and modeling, by: (i) evaluating how AI-produced, human-authored or hybrid human-AI news content differ in informativeness, bias, and engagement (Zhou et al., 2025b; Ziv et al., 2025); (ii) designing debiasing and robustness mechanisms that differentiate between synthetic and human sources, for example via metadata-aware encoders and provenance-aware ranking (Khan et al., 2025; Kehkashan et al., 2025); (iii) integrating transparency signals, such as AI content labels or verifiable content credentials into recommendation objectives to support fairness, accountability, and trust; and (iv) assessing how AI-generated content interacts with user perceptions, misinformation propagation, and democratic discourse (Park and Nan, 2025). These directions respond to a rapidly evolving news landscape in which AI increasingly acts as an active content producer, generating articles, summaries, and personalized narratives, while escalating the risks of misinformation, bias amplification, and reduced transparency. Addressing these developments requires methods that ensure factual accuracy, source reliability, and equitable exposure, while responsibly harnessing generative models to enhance personalization, context-aware recommendation, and richer user experiences.

Implications Beyond the News Domain. While this thesis focuses on multifaceted neural news recommendation, many of the challenges it addresses are not unique to the news domain. These include the role of common components in driving overall recommender behavior, inherent biases in recommender systems and biased exposure, diversity-aware recommendation, and the tension between relevance and societal objectives. An important future direction therefore lies in translating the conceptual insights and technical mechanism developed here to other recommendation domains.

The Need for Multi-Criteria and Rigorous Evaluation of Recommenders and Their Components. Systematically and fairly analyzing the role and interaction of architectural components beyond downstream accuracy (cf. Chapters 5 and 6) resonates with broader calls in recommender systems

[1] https://contentauthenticity.org/

research to scrutinize how design choices shape overall system behavior. For example, work by Anelli et al. (2021a) re-evaluating collaborative filtering against classic baselines shows that focusing solely on accuracy can obscure important differences in diversity, novelty, item coverage, and bias effects between competing architectures. Related studies further demonstrate that, under rigorous and transparent evaluation protocols, conceptually simpler methods can offer competitive, and sometimes, more robust performance compared to more complex models (Ferrari Dacrema et al., 2019, 2021; Anelli et al., 2021a; Latifi and Jannach, 2022). Similarly, systematic reviews of recommender evaluation practice find that accuracy-centric metrics dominate empirical comparisons, while beyond-accuracy dimensions such as serendipity, fairness, or coverage, receive far less attention, prompting calls for more comprehensive evaluation frameworks (Bauer et al., 2024; Knees et al., 2024). Recent work explicitly argues for moving beyond accuracy-centric evaluation, showing that design choices across multiple components of the recommendation pipeline systematically interact with these higher-order objectives (Duricic et al., 2023). Research on benchmarking further emphasizes that fair and meaningful model comparison requires standardized, multi-criteria evaluation protocols that account for architectural and optimization factors (Sun et al., 2020).

Together, these examples underscore the broader relevance of treating recommender systems as collections of interacting components and evaluating them from multiple complementary perspectives. While the specific techniques developed in this thesis may not directly transfer to other domains, the lessons learned – that simpler architectures can outperform more complex ones under certain criteria, that representation quality matters as much as aggregate prediction scores, and that responsible evaluation must extend beyond accuracy-only metrics – can inspire analogous analyses across other domains, including e-commerce, entertainment, scholarly or social content recommendation.

Modularity and Multi-Aspect Customization as Design Paradigm for Responsible Recommenders. The societal effects of algorithmic content curation discussed in this thesis extend well beyond the news domain. Across a wide range of recommender systems, algorithmic ranking and personalization has been shown to exhibit systemic biases and diversity limitations that shape user exposure in meaningful ways. In media recommendation, for example, popularity bias has been found to disproportionately disadvantage female users in music recommendation, amplifying gender disparities in exposure and user experience (Melchiorre et al., 2021; Lesota et al., 2021). Similarly, studies of movie recommendation contexts reveal algorithm-induced stereotypes and disparate impact on atypical or minority users, highlighting that biased exposure and representational imbalance are pervasive phenomena across domains (Ekstrand et al., 2018; Ahn and Lin, 2024). Beyond media consumption, recommender systems in the scholarly domain – which suggest articles, collaborators, or venues to researchers – are increasingly recognized as susceptible to structural biases such as popularity and exposure effects, as well as demographic skew, influencing the visibility and perceived impact of researchers across gender, career stage, or institutional background (Färber et al., 2023). In job and recruitment settings, algorithmic bias can have direct material consequences: empirical audits show that female and male applicants may receive systematically different recommendations in terms of salary levels and stereotypically gendered roles, affecting access to opportunities (Rus et al., 2022; Zhang and Kuhn, 2024; Bied et al., 2023). An emerging body of work on cultural recommendation further highlights that systems curating cultural goods shape collective cultural experience and, when optimized solely for personalized utility, may marginalize underrepresented groups or reinforce dominant cultural norms in domains such as entertainment or e-commerce (Ferraro et al., 2022; Moradi et al., 2025). Across these settings, diversity is inherently multi-dimensional and normative, encompassing not only item-level diversity, but also demographic, cultural, and societal representation, and raising fundamental questions about which forms of visibility and exposure recommender systems should promote, and for whom.

Within the news domain, the applicability of the proposed MANNeR framework to frame-based normative diversity has already been demonstrated on additional Portuguese and Danish datasets (Dattawad et al., 2025), showing that multi-aspect ranking and user-controllable trade-offs between

relevance, diversification, and personalization generalize across languages, datasets, and perspective dimensions. Beyond news, the recurring need to evaluate and actively control exposure, diversity, and fairness closely parallels the motivations of this thesis (cf. § 1.1.2) and suggests that frameworks like MANNeR could be adapted to support normative diversity and equitable exposure in other recommendation contexts. From this perspective, the modular and multi-aspect design of MANNeR points toward a more general design paradigm for responsible recommender systems, in which relevance is treated as one objective among several, and trade-offs are made explicit, transparent, and adaptable to different stakeholders and domains. Future work could investigate how such modular, aspect-based approaches transfer to other domains, such as scholarly, cultural, or job recommendation, where notions of diversity, fairness, and exposure take different but equally consequential forms.

Multilinguality as Driver of Linguistic Inclusivity in Recommendation. While this thesis addresses multilinguality in the context of neural news recommendation, the challenges it reveals are not specific to news. Across many areas of AI, such as recommendation, speech recognition, conversational agents, and information access systems, language profoundly shapes which items users can access, how content is represented, and how effectively personalization models serve different user populations. Yet, much like in news recommendation, research in these domains remains largely dominated by monolingual benchmarks and models optimized for resource-rich languages, implicitly assuming homogeneous linguistic behavior. The findings of this thesis highlight several insights with broader relevance for addressing this gap. First, multilingual evaluation resources such as xMIND expose performance disparities and failure modes that remain hidden under monolingual evaluation, particularly for speakers of low-resource languages and multilingual users. Second, the strong gains achieved by domain-adapted but frozen multilingual representations, as demonstrated with NaSE, underscore that robust and transferable semantic representations can be more effective than computationally expensive task-specific fine-tuning in data-scarce or low-resource settings, conditions that frequently arise beyond news recommendation.

From a societal perspective, these linguistic performance gaps directly translate into unequal exposure and access to information, making multilinguality not only a technical challenge but also a fairness concern, as users' language choices can systematically influence what content and opportunities they are recommended. Related work in other recommendation domains additionally supports these observations. For instance, a study on multilingual prompts in LLM-based recommendation shows that supporting non-English languages can reduce performance gaps and improve recommendation quality across linguistic groups (Ozsoy, 2024). Similarly, efforts in bilingual conversational recommendation benchmarks demonstrate that explicitly modeling cross-lingual user-item interactions leads to better generalization for multilingual users (Liu et al., 2021c). Overall, these results suggest that advancing multilingual recommender systems across domains requires more inclusive evaluation practices and datasets that reflect linguistic diversity, as well as modeling approaches that prioritize representation robustness, scalability, and cross-lingual generalization. The resources and methodologies introduced in Chapters 8 and 9, such as linguistically and geographically diverse datasets, rigorous benchmarking, and domain-specialized multilingual encoders, can thus inform the design of more equitable and linguistically inclusive recommender systems across a wide range of multilingual application domains.

As a scientific community developing these technologies, we bear responsibility not only for improving the effectiveness of news recommender systems, but also for ensuring that they are fair, inclusive, and robust across languages and contexts. We hope that our thesis inspires and catalyzes follow-up work towards achieving this goal.

BIBLIOGRAPHY

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467*.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, Andre N. Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin A. Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire M. Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News Topic Classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Yongsu Ahn and Yu-ru Lin. 2024. Break Out of a Pigeonhole: A Unified Framework for Examining

Miscalibration, Bias, and Stereotype in Recommender Systems. *ACM Transactions on Intelligent Systems and Technology*, 15(4):1–20.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mehwish Alam, Andreea Iana, Alexander Grote, Katharina Ludwig, Philipp Müller, and Heiko Paulheim. 2022. Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 448–457, New York, NY, USA. Association for Computing Machinery.

Amanda Alencar and Mark Deuze. 2017. News for Assimilation or Integration? Examining the Functions of News in Shaping Acculturation Experiences of Immigrants in the Netherlands and Spain. *European Journal of Communication*, 32(2):151–166.

Manal A Alshehri and Xiangliang Zhang. 2022. Generative Adversarial Zero-Shot Learning for Cold-Start News Recommendation. In *Proceedings of the 31st ACM International Conference on Information Knowledge Management*, pages 26–36.

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345.

Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, and Claudio Pomo. 2021a. Reenvisioning the comparison between neural collaborative filtering and matrix factorization. In *Proceedings of the 15th ACM conference on recommender systems*, pages 521–529.

Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021b. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2405–2414.

Abolfazl Ansari, Delvin Ce Zhang, Nafis Irtiza Tripto, and Dongwon Lee. 2025. Echoes of Automation: The Increasing Use of LLMs in Newsmaking. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 3–13. Springer.

Andreas Argyriou, Miguel González-Fierro, and Le Zhang. 2020. Microsoft Recommenders: Best Practices for Production-Ready Recommendation Systems. In *Companion Proceedings of the Web Conference 2020*, pages 50–51.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *arXiv preprint arXiv:1907.05019*.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear Algebraic Structure of Word Senses, with Applications to Polysemy . *Transactions of the Association for Computational Linguistics*, 6:483–495.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-beat Baseline for Sentence Embeddings. In *International conference on learning representations*.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. MorphBPE: A Morpho-Aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies. *arXiv preprint arXiv:2502.00894*.

Yonata Andrelo Asikin and Wolfgang Wörndl. 2014. Stories Around You: Location-based Serendipitous Recommendation of News Articles. In *UMAP Workshops*.

Igor LR Azevedo, Toyotaro Suzumura, and Yuichiro Yasui. 2024. Popular News Always Compete for the User's Attention!POPK: Mitigating Popularity Bias via a Temporal-Counterfactual. *arXiv preprint arXiv:2407.09939*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450*.

Christian Baden and Nina Springer. 2017. Conceptualizing Viewpoint Diversity in News Discourse. *Journalism*, 18(2):176–194.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Xiao Bai, B Barla Cambazoglu, Francesco Gullo, Amin Mantrach, and Fabrizio Silvestri. 2017. Exploiting Search History of Users for News Personalization. *Information Sciences*, 385:125–137.

C.E. Baker. 2006. *Media Concentration and Democracy: Why Ownership Matters*. Communication, Society and Politics. Cambridge University Press.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*, 348(6239):1130–1132.

Jack M Balkin. 2017. Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *UCDL rev.*, 51:1149.

Jack Bandy and Nicholas Diakopoulos. 2020. Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 36–47.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

Pablo Barberá. 2020. *Social Media, Echo Chambers, and Political Polarization*, SSRC Anxieties of Democracy, pages 34–55. Cambridge University Press.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022a. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022b. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266.

Larry M Bartels. 1993. Messages Received: The Political Impact of Media Exposure. *American political science review*, 87(2):267–285.

Marco Bastos, Dan Mercea, and Andrea Baronchelli. 2018. The Geographic Embedding of Online Echo Chambers: Evidence from the Brexit Campaign. *PloS one*, 13(11):e0206841.

Christine Bauer, Eva Zangerle, and Alan Said. 2024. Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives. *ACM Transactions on Recommender Systems*, 2(1):1–31.

BBC. 2022. "Inform, educate, entertain... and recommend?" - Public service recommendations. *BBC Blog*. [Online; accessed 12-August-2025].

Michael A Beam. 2014. Automating the News: How Personalized News Recommender System Design Choices Impact News Reception. *Communication Research*, 41(8):1019–1041.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In *First Conference on Language Modeling*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Emily Bender. 2019. The BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*, 14.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of machine learning research*, 3(Feb):1137–1155.

James Bennett, Stan Lanning, et al. 2007. The Netflix Prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, page 35. New York.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems*, 24.

Abraham Bernstein, Claes de Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A. Beam, Marc P. Hauer, Lucien Heitz, Pascal Jürgens, Christian Katzenbach, Benjamin Kille, Beate Klimkiewicz, Wiebke Loosen, Judith Moeller, Goran Radanovic, Guy Shani, Nava Tintarev, Suzanne Tolmeijer, Wouter van Atteveldt, Sanne Vrijenhoek, and Theresa Zueger. 2021. Diversity in News Recommendation (Dagstuhl Perspectives Workshop 19482). *Dagstuhl Manifestos*, 9(1):43–61.

Guillaume Bied, Christophe Gaillac, Morgane Hoffmann, Philippe Caillou, Bruno Crépon, Solal Nathan, and Michele Sebag. 2023. Fairness in Job Recommendations: Estimating, Explaining, and Reducing Gender Gaps. In *AEQUITAS 2023-First AEQUITAS Workshop on Fairness and Bias in AI| co-located with ECAI 2023*, volume 3523. CEUR-WS. org.

Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. Software available from wandb.com.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information . *Transactions of the association for computational linguistics*, 5:135–146.

Vincenzo Bonnici. 2020. Kullback-Leibler Divergence between Quantum Distributions, and Its Upper-bound. *arXiv preprint arXiv:2008.05932*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. *Advances in Neural Information Processing Systems*, 26.

Andrei Boutyline and Robb Willer. 2017. The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks. *Political psychology*, 38(3):551–569.

Çiğdem Bozdağ and Yasemin Karakasoğlu. 2024. Multilingual Media Repertoires of Young People in the Migration Society: A Plea for a Language and Culture-aware Approach to Media Education. *Global Studies of Childhood*, 14(4):448–461.

Engin Bozdag. 2013. Bias in Algorithmic Filtering and Personalization. *Ethics and information technology*, 15(3):209–227.

Jonathan Bright. 2018. Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism. *Journal of computer-mediated communication*, 23(1):17–33.

Emma Brocken, Aron Hartveld, Emma de Koning, Thomas van Noort, Frederik Hogenboom, Flavius Frasincar, and Tarmo Robal. 2019. Bing-CF-IDF+: A Semantics-Driven News Recommender System. In *International Conference on Advanced Information Systems Engineering*, pages 32–47. Springer.

Andrei Z Broder. 1997. On the Resemblance and Containment of Documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Davis Brown, Charles Godfrey, Nicholas Konz, Jonathan Tu, and Henry Kvinge. 2023. Understanding the Inner-workings of Language Models Through Representation Dissimilarity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6543–6558.

Lawrence Bunnell, Kweku-Muata Osei-Bryson, and Victoria Y Yoon. 2020. FinPathlight: Framework for an Multiagent Recommender System Designed to Increase Consumer Financial Capability. *Decision Support Systems*, 134:113306.

Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Robin Burke. 2017. Multisided Fairness for Recommendation. *arXiv preprint arXiv:1707.00093*.

Robin Burke and Michael Ekstrand. 2025. Conducting Recommender Systems User Studies Using POPROX. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 1–2.

Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE transactions on knowledge and data engineering*, 30(9):1616–1637.

Michel Capelle, Flavius Frasincar, Marnix Moerland, and Frederik Hogenboom. 2012. Semantics-based News Recommendation. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, pages 1–9.

Michel Capelle, Marnix Moerland, Frederik Hogenboom, Flavius Frasincar, and Damir Vandic. 2015. Bing-sf-idf+: A hybrid semantics-driven news recommender. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 732–739.

Laura Caspari, Kanishka Ghosh Dastidar, Saber Zerhoudi, Jelena Mitrovic, and Michael Granitzer. 2024. Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems. In *Proceedings of the Workshop Information Retrieval's Role in RAG Systems (IR-RAG 2024) co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*.

Pablo Castells, Neil Hurley, and Saul Vargas. 2021. Novelty and Diversity in Recommender Systems. In *Recommender systems handbook*, pages 603–646. Springer.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kinjal Chaudhari and Ankit Thakkar. 2020. A Comprehensive Survey on Travel Recommender Systems. *Archives of computational methods in engineering*, 27(5):1545–1571.

Cheng Chen, Xiangwu Meng, Zhenghua Xu, and Thomas Lukasiewicz. 2017. Location-aware Personalized News Recommendation with Deep Semantic Analysis. *IEEE Access*, 5:1624–1638.

Hao Chen, Zefan Wang, Feiran Huang, Xiao Huang, Yue Xu, Yishi Lin, Peng He, and Zhoujun Li. 2022. Generative Adversarial Framework for Cold-Start Item Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2565–2571.

Jianlyu Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, and Zheng Liu. 2025a. AIR-Bench: Automated Heterogeneous Information Retrieval Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19991–20022.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023a. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.*, 41(3).

Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior Sequence Transformer for E-commerce Recommendation in Alibaba. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, New York, NY, USA. Association for Computing Machinery.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending Context Window of Large Language Models via Positional Interpolation. *arXiv preprint arXiv:2306.15595*.

Weixin Chen, Li Chen, Yongxin Ni, and Yuhan Zhao. 2025b. Causality-Inspired Fair Representation Learning for Multimodal Recommendation. *ACM Transactions on Information Systems*, 43(6):1–29.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3576–3588.

Jaeho Cho, Saifuddin Ahmed, Martin Hilbert, Billy Liu, and Jonathan Luu. 2020. Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization. *Journal of Broadcasting & Electronic Media*, 64(2):150–172.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1724. Association for Computational Linguistics.

Sungmin Cho, Hongjun Lim, Keunchan Park, Sungjoo Yoo, and Eunhyeok Park. 2021. On the Overlooked Significance of Underutilized Contextual Features in Recent News Recommendation Models. *arXiv preprint arXiv:2112.14370*.

Seonghwan Choi, Hyeondey Kim, and Manjun Gim. 2022. Do Not Read the Same News! Enhancing Diversity and Personalization of News Recommendation. In *Companion Proceedings of the Web Conference 2022*, pages 1211–1215.

Qianfeng Chu, Gongshen Liu, Huanrong Sun, and Cheng Zhou. 2019. Next News Recommendation via Knowledge-Aware Sequential Model. In *China National Conference on Chinese Computational Linguistics*, pages 221–232. Springer.

Wei Chu and Seung-Taek Park. 2009. Personalized Recommendation on Dynamic Content using Predictive Bilinear Models. In *Proceedings of the 18th International Conference on World Wide Web*, pages 691–700.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining. In *The Eleventh International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, and Matthew Sartin. 1999. Combining Content-based and Collaborative filters in an Online Newspaper. In *Proceedings of ACM SIGIR workshop on recommender systems*, volume 60, pages 1853–1870. Berkley, CA, USA.

Simone Conia, Min Li, Daniel Lee, Umar Minhas, Ihab Ilyas, and Yunyao Li. 2023. Increasing Coverage and Precision of Textual Information in Multilingual Knowledge Graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1612–1634.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7059–7069.

Emile Contal and Garrin McGoldrick. 2024. RAGSys: Item-Cold-Start Recommender as RAG System. In *Proceedings of the Workshop Information Retrieval's Role in RAG Systems (IR-RAG 2024)co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*, volume 3784.

Francesco Corsini and Martha A Larson. 2016. CLEF NewsREEL 2016: Image-based Recommendation. In *CLEF 2016: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, pages 593–605. CEUR.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, page 191–198, New York, NY, USA. Association for Computing Machinery.

Arthur Da Costa, Eduardo Fressato, Fernando Neto, Marcelo Manzato, and Ricardo Campello. 2018. Case Recommender: A Flexible and Extensible Python Framework for Recommender Systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 494–495.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Agnese Daffara, Sourabh Dattawad, Sebastian Padó, and Tanise Ceron. 2025. Generalizability of Media Frames: Corpus Creation and Analysis across Countries. *arXiv preprint arXiv:2506.16337*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge Neurons in Pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Pranav Dandekar, Ashish Goel, and David T Lee. 2013. Biased Assimilation, Homophily, and the Dynamics of Polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626.

Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 271–280, New York, NY, USA. Association for Computing Machinery.

Sourabh Dattawad, Agnese Daffara, and Tanise Ceron. 2025. Leveraging Media Frames to Improve Normative Diversity in News Recommendations. In *INRA 2025: Proceedings of the 13th International Workshop on News Recommendation and Analytics co-located with the 2025 ACM Conference on Recommender Systems (RecSys 2025)*, volume Vol-4056.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual Autoregressive Entity Linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Emma de Koning, Frederik Hogenboom, and Flavius Frasincar. 2018. News recommendation with cf-idf+. In *International Conference on Advanced Information Systems Engineering*, pages 170–184. Springer.

Marco De Nadai, Andreas Damianou, and Mounia Lalmas. 2025. Describe What You See with Multimodal Large Language Models to Enhance Video Recommendations. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 1159–1163.

Marco De Nadai, Francesco Fabbri, Paul Gigioli, Alice Wang, Ang Li, Fabrizio Silvestri, Laura Kim, Shawn Lin, Vladan Radosavljevic, Sandeep Ghael, David Nyhan, Hugues Bouchard, Mounia Lalmas, and Andreas Damianou. 2024. Personalized Audiobook Recommendations at Spotify Through Graph Neural Networks. In *Companion Proceedings of the ACM Web Conference 2024*, page 403–412, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, Rene Vidal, Maheswaran Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, et al. 2024a. Recommendation with Generative Models. *arXiv preprint arXiv:2409.15173*.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2024b. Fairness in Recommender Systems: Research Landscape and Future Directions. *User Modeling and User-Adapted Interaction*, 34(1):59–108.

Derivation. 2022. Understanding the Impact of Digital Language Support. [Online; accessed 26-August-2025].

Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. 2022. TorchMetrics - Measuring Reproducibility in PyTorch.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ralf Dewenter, Melissa Linder, and Tobias Thomas. 2019. Can Media Drive the Electorate? The Impact of Media Coverage on Voting Intentions. *European Journal of Political Economy*, 58:245–261.

Dario Di Palma. 2023. Retrieval-augmented Recommender System: Enhancing Recommender Systems with Large Language Models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1369–1373.

Wolfgang Donsbach and Cornelia Mothes. 2013. The Dissonant Self: Contributions from Dissonance Theory to a New Agenda for Studying Political Communication. *Annals of the International Communication Association*, 36(1):3–44.

Doychin Doychev, Rachael Rafter, Aonghus Lawlor, and Barry Smyth. 2015. News Recommenders: Real-Time, Real-Life Experiences. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 337–342. Springer.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-Level Multimodal and Language-Agnostic Representations. *arXiv e-prints*, pages arXiv–2308.

Tomislav Duricic, Dominik Kowald, Emanuel Lacic, and Elisabeth Lex. 2023. Beyond-accuracy: A review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. *Frontiers in big data*, 6:1251072.

Cynthia Dwork. 2006. Differential Privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

David M. Eberhard, Gary F. Simons, and Charles D. (eds.) Fennig. 2025. Ethnologue: Languages of the World. [Online; accessed 26-August-2025].

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Árni Már Einarsson, R Helles, and S Lomborg. 2025. Algorithmic agenda-setting: the subtle effects of news recommender systems on political agendas in the Danish 2022 general election. *Information, Communication & Society*, 28(2):218–238.

Michael D Ekstrand. 2020. LensKit for Python: Next-Generation Software for Recommender Systems Experiments. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2999–3006.

Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. *Fairness in Recommender Systems*, pages 679–707. Springer US, New York, NY.

Michael D Ekstrand, Michael Ludwig, Joseph A Konstan, and John T Riedl. 2011. Rethinking the Recommender Research Ecosystem: Reproducibility, Openness, and Lenskit. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 133–140.

Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Conference on fairness, accountability and transparency*, pages 172–186. PMLR.

Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, et al. 2022. Towards Responsible Media Recommendation. *AI and Ethics*, 2(1):103–114.

Jonas Elis. 2025. How Do Immigrant-Origin and Native Voters Consume Political News Media During a National Election Campaign? *Politische Vierteljahresschrift*, pages 1–26.

Elena Viorica Epure, Benjamin Kille, Jon Espen Ingvaldsen, Rebecca Deneckere, Camille Salinesi, and Sahin Albayrak. 2017. Recommending Personalized News in Short User Sessions. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 121–129.

Sarah Eskens. 2022. Regulating News Recommender Systems in Light of the Rule of Law: Recommender Systems in the Online News Sector. *Verfassungsblog*. [Online; accessed 07-November-2025].

Maryam Etemadi, Sepideh Bazzaz Abkenar, Ahmad Ahmadzadeh, Mostafa Haghi Kashani, Parvaneh Asghari, Mohammad Akbari, and Ebrahim Mahdipour. 2023. A Systematic Review of Healthcare Recommender Systems: Open Issues, Challenges, and Techniques. *Expert Systems with Applications*, 213:118823.

European Commission. 2019. Ethics Guidelines for Trustworthy AI. `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`. [Online; accessed 07-November-2025].

European Parliament, Council of the European Union, European Commission. 2023. European Declaration on Digital Rights and Principles for the Digital Decade 2023/C 23/01. Official Journal of the European Union. [Online; accessed 12-August-2025].

Eurostat. 2024. Digital economy and society statistics - households and individuals. [Online; accessed 12-August-2025].

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3007–3014.

Michael Färber, Melissa Coutinho, and Shuzhou Yuan. 2023. Biases in Scholarly Recommender Systems: Impact, Prevalence, and Mitigation. *Scientometrics*, 128(5):2703–2736.

Chong Feng, Muzammil Khan, Arif Ur Rahman, and Arshad Ahmad. 2020. News Recommendation Systems - Accomplishments, Challenges Future Directions. *IEEE Access*, 8:16702–16725.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Syeda Nyma Ferdous and Muhammad Masroor Ali. 2017. A Semantic Content Based Recommendation System for Cross-lingual News. In *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 1–6. IEEE.

Miriam Fernandez, Alejandro Bellogín, and Iván Cantador. 2024. Analysing the Effect of Recommendation Algorithms on the Spread of Misinformation. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, page 159–169, New York, NY, USA. Association for Computing Machinery.

Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. 2012. Cross-domain Recommender Systems: A survey of the State of the Art. In *Spanish Conference on Information Retrieval*, volume 24. sn.

Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–49.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proceedings of the 13th ACM conference on recommender systems*, pages 101–109.

Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. 2022. Measuring Commonality in Recommendation of Cultural Content: Recommender Systems to Enhance Cultural Citizenship. In *Proceedings of the 16th ACM conference on recommender systems*, pages 567–572.

John Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32.

Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320.

Richard Fletcher and Rasmus Kleis Nielsen. 2017. Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication. *Journal of Communication*, 67(4):476–498.

Blaž Fortuna, Carolina Fortuna, and Dunja Mladenić. 2010. Real-Time News Recommender System. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 583–586. Springer.

Jonathan L Freedman and David O Sears. 1965. Selective exposure. In *Advances in experimental social psychology*, volume 2, pages 57–97. Elsevier.

Matthew Freestone and Shubhra Kanti Karmaker Santu. 2024. Word Embeddings Revisited: Do LLMs Offer Something New? *arXiv preprint arXiv:2402.11094*.

P Moreira Gabriel De Souza, Dietmar Jannach, and Adilson Marques Da Cunha. 2019. Contextual Hybrid Session-based News Recommendation with Recurrent Neural Networks. *IEEE Access*, 7:169185–169203.

Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. 2004. Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. In *Proceedings of the 13th international conference on World Wide Web*, pages 482–490.

Philip Gage. 1994. A New Algorithm for Data Compression. *The C Users Journal*, 12(2):23–38.

Petra Galuscáková, Romain Deveaud, Gabriela González Sáez, Philippe Mulhem, Lorraine Goeuriot, Florina Piroi, and Martin Popel. 2023. LongEval-Retrieval: French-English Dynamic Test Collection for Continuous Web Search Evaluation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3086–3094.

Petra Galuščáková, Douglas W Oard, and Suraj Nair. 2021. Cross-language Information Retrieval. *arXiv preprint arXiv:2111.05988*.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting Political Bias in News Articles Using Headline Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84.

Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl. 2024. A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 380–390.

Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. MyMediaLite: A Free Recommender System Library. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 305–308.

Jie Gao, Xin Xin, Junshuai Liu, Rui Wang, Jing Lu, Biao Li, Xin Fan, and Ping Guo. 2018a. Fine-grained Deep Knowledge-aware Network for News Recommendation with Self-attention. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 81–88. IEEE.

Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018b. To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–16.

Shen Gao, Jiabao Fang, Quan Tu, Zhitao Yao, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2024. Generative News Recommendation. In *Proceedings of the ACM Web Conference 2024*, pages 3444–3453.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alba Táboas García, Piotr Przybyła, and Leo Wanner. 2025. Exploring Morphology-Aware Tokenization: A Case Study on Spanish Language Modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30493–30506.

Florent Garcin, Christos Dimitrakakis, and Boi Faltings. 2013. Personalized News Recommendation with Context Trees. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 105–112.

Florent Garcin, Kai Zhou, Boi Faltings, and Vincent Schickel. 2012. Personalized News Recommendation Based on Collaborative Filtering. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 437–441. IEEE.

Angel Luis Garrido, María G Buey, Sergio Ilarri, Igor Fũrstner, and Livia Szedmina. 2015. KGNR: A Knowledge-based Geographical News Recommender. In *2015 IEEE 13th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 195–198. IEEE.

Suyu Ge, Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. Graph Enhanced Representation Learning for News Recommendation. In *Proceedings of the Web Conference 2020*, pages 2863–2869.

Anatole Gershman, Travis Wolfe, Eugene Fink, and Jaime G Carbonell. 2011. News Personalization Using Support Vector Machines.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Sahin Cem Geyik, Qi Guo, Bo Hu, Cagri Ozcaglar, Ketan Thakkar, Xianren Wu, and Krishnaram Kenthapadi. 2018. Talent Search and Recommendation Systems at LinkedIn: Practical Challenges and Lessons Learned. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 1353–1354, New York, NY, USA. Association for Computing Machinery.

Alireza Gharahighehi and Celine Vens. 2023. Diversification in Session-based News Recommender Systems. *Personal and Ubiquitous Computing*, 27(1):5–15.

Alireza Gharahighehi, Celine Vens, and Konstantinos Pliakos. 2021. Fair Multi-stakeholder News Recommender System with Hypergraph Ranking. *Information Processing & Management*, 58(5):102663.

Iris Glassberg, Yael Brender Ilan, and Moti Zwilling. 2025. The Key Role of Design and Transparency in Enhancing Trust in AI-powered Digital Agents. *Journal of Innovation & Knowledge*, 10(5):100770.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett, et al. 2018. Fake News vs Satire: A Dataset and Analysis. In *Proceedings of the 10th ACM Conference on Web Science*, pages 17–21.

Yoav Goldberg and Omer Levy. 2014. Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Carlos A Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19.

Shansan Gong and Kenny Q Zhu. 2022. Positive, Negative and Neutral: Modeling Implicit Feedback in Session-based News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1185–1195.

Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. Asymmetric Ideological Segregation in Exposure to Political News on Facebook. *Science*, 381(6656):392–398.

Cristos Goodrow. 2025. On YouTube's Recommendation System. *YouTube Official Blog*. [Online; accessed 11-August-2025].

Frank Goossen, Wouter IJntema, Flavius Frasincar, Frederik Hogenboom, and Uzay Kaymak. 2011. News Personalization Using the CF-IDF Semantic Recommender. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, pages 1–12.

Jyotirmoy Gope and Sanjay Kumar Jain. 2017. A Survey on Solving Cold Start Problem in Recommender Systems. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 133–138. IEEE.

Scott Graham, Jun-Ki Min, and Tao Wu. 2019. Microsoft recommenders: Tools to accelerate developing recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 542–543.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794*.

François Grosjean. 2010. *Bilingual: Life and Reality*. Harvard University Press.

Thomas R Gruber. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing? *International journal of human-computer studies*, 43(5-6):907–928.

Wanrong Gu, Shoubin Dong, and Mingquan Chen. 2016. Personalized news recommendation based on articles chain building. *Neural Computing and Applications*, 27(5):1263–1272.

Wanrong Gu, Shoubin Dong, Zhizhao Zeng, and Jinchao He. 2014. An Effective News Recommendation Method for Microblog User. *The Scientific World Journal*, 2014(1):907515.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1627–1632, Marseille, France. European Language Resources Association.

Jon Atle Gulla, Arne Dag Fidjestøl, Xiaomeng Su, and Humberto Castejon. 2014. Implicit User Profiling in News Recommender Systems. In *International Conference on Web Information Systems and Technologies*, volume 2, pages 185–192. SCITEPRESS.

Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The Adressa Dataset for News Recommendation. In *Proceedings of the International Conference on Web Intelligence*, pages 1042–1048.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2020a. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *International Conference on Learning Representations*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2020b. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *International Conference on Learning Representations*.

Guibing Guo, Jie Zhang, Zhu Sun, and Neil Yorke-Smith. 2015. Librec: A Java Library for Recommender Systems. In *UMAP workshops*, volume 4, pages 38–45. Citeseer.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2020a. A Deep Look into Neural Ranking Models for Information Retrieval. *Information Processing & Management*, 57(6):102067.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020b. A Survey on Knowledge Graph-Based Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568.

Taicheng Guo, Lu Yu, Basem Shihada, and Xiangliang Zhang. 2023. Few-shot News Recommendation via Cross-lingual Transfer. In *Proceedings of the ACM Web Conference 2023*, pages 1130–1140.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304. JMLR Workshop and Conference Proceedings.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation . *Computational Linguistics*, 48(3):673–732.

Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the Filter Bubble? Effects of Personalization on the Diversity of Google News. *Digital journalism*, 6(3):330–343.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated Identification of Media Bias in News Articles: An Interdisciplinary Literature Review. *International Journal on Digital Libraries*, 20(4):391–415.

David L Hamilton and Robert K Gifford. 1976. Illusory Correlation in Interpersonal Perception: A Cognitive Basis of Stereotypic Judgments. *Journal of Experimental Social Psychology*, 12(4):392–407.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems*, 30.

Jihun Hamm and Daniel D Lee. 2008. Grassmann Discriminant Analysis: A Unifying View on Subspace-based Learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 376–383.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. glottolog/glottolog: Glottolog database 4.4.

Songhao Han, Wei Huang, and Xiaotian Luan. 2022. VLSNR:Vision-Linguistics Coordination Time Sequence-aware News Recommendation. *arXiv preprint arXiv:2210.02946*.

Songqiao Han, Hailiang Huang, and Jiangwei Liu. 2021. Neural News Recommendation with Event Extraction. *arXiv preprint arXiv:2111.05068*.

Mahboobeh Harandi and Jon Atle Gulla. 2015. Survey of user profiling in news recommender systems. In *INRA@ RecSys*, pages 20–26.

F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Zellig S Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Syed Zain Ul Hassan, Muhammad Rafi, and Jaroslav Frnda. 2024. GCZRec: Generative Collaborative Zero-shot Framework for Cold Start News Recommendation. *IEEE Access*, 12:16610–16620.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Nicolas Heist, Sven Hertling, Daniel Ringler, and Heiko Paulheim. 2020. Knowledge Graphs on the Web – An Overview. *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, pages 3–22.

Lucien Heitz, Runze Li, Oana Inel, and Abraham Bernstein. 2025. Informfully Recommenders-Reproducibility Framework for Diversity-aware Intra-session Recommendations. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 792–801.

Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. 2022. Benefits of Diverse News Recommendations for Democracy: A User Study. *Digital Journalism*, 10(10):1710–1730.

Lucien Heitz, Nicolas Mattis, Oana Inel, and Wouter van Atteveldt. 2024. IDEA–Informfully Dataset with Enhanced Attributes. In *Proceedings of the Second Workshop on the Normative Design and Evaluation of Recommender Systems (NORMalize 2024) co-located with the 18th ACM Conference on Recommender Systems (RecSys 2024)*, pages 1–12. CEUR-WS (CEUR Workshop Proceedings).

Natali Helberger. 2021. On the Democratic Role of News Recommenders. In *Algorithms, Automation, and News*, pages 14–33. Routledge.

Karl Higley, Robin Burke, Michael D Ekstrand, and Bart P Knijnenburg. 2025. What News Recommendation Research Did (But Mostly Didn't) Teach Us About Building A News Recommender. In *Beyond Algorithms: Reclaiming the Interdisciplinary Roots of Recommender Systems Workshop (BEYOND 2025), September 26th, 2025, co-located with the 19th ACM Recommender Systems Conference, Prague, Czech Republic*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

Frederik Hogenboom, Michel Capelle, and Marnix Moerland. 2013. News Recommendation Using Semantics with the Bing-SF-IDF Approach. In *International Conference on Conceptual Modeling*, pages 160–169. Springer.

Benjamin Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Cheng-Kang Hsieh, Longqi Yang, Honghao Wei, Mor Naaman, and Deborah Estrin. 2016. Immersive Recommendation: News and Event Recommendations Using Personal Digital Traces. In *Proceedings of the 25th International Conference on World Wide Web*, pages 51–62.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 1(2):3.

Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. 2020a. Graph Neural News Recommendation with Long-term and Short-term Interest Modeling. *Information Processing & Management*, 57(2):102142.

Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou. 2020b. Graph Neural News Recommendation with Unsupervised Preference Disentanglement. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4255–4264.

Feiran Huang, Yuanchen Bei, Zhenghang Yang, Junyi Jiang, Hao Chen, Qijie Shen, Senzhang Wang, Fakhri Karray, and Philip S Yu. 2025a. Large Language Model Simulator for Cold-Start Recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 261–270.

Feiran Huang, Zefan Wang, Xiao Huang, Yufeng Qian, Zhetao Li, and Hao Chen. 2023. Aligning Distillation For Cold-start Item Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1147–1157.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025b. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Sirui Huang, Qian Li, Haoran Yang, Dianer Yu, Qing Li, and Guandong Xu. 2025c. Causal Time-aware News Recommendations with Large Language Models. *ACM Transactions on Information Systems*, 43(6):1–25.

Yongxin Huang, Kexin Wang, Goran Glavaš, and Iryna Gurevych. 2025d. Modular Sentence Encoders: Separating Language Specialization from Cross-Lingual Alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2167–2187.

Nicolas Hug. 2020. Surprise: A Python Library for Recommender Systems. *Journal of Open Source Software*, 5(52):2174.

Chia-Chien Hung, Lukas Lange, and Jannik Strötgen. 2023. TADA: Efficient Task-Agnostic Domain Adaptation for Transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 487–503, Toronto, Canada. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Simone Ponzetto, and Goran Glavaš. 2022. DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904, Dublin, Ireland. Association for Computational Linguistics.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic Amplification of Politics on Twitter. *Proceedings of the national academy of sciences*, 119(1):e2025334119.

Clayton Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Andreea Iana, Mehwish Alam, Alexander Grote, Nevena Nikolajevic, Katharina Ludwig, Philipp Müller, Christof Weinhardt, and Heiko Paulheim. 2022a. NeMig - A Bilingual News Collection and Knowledge Graph about Migration.

Andreea Iana, Mehwish Alam, and Heiko Paulheim. 2024. A Survey on Knowledge-aware News Recommender Systems. *Semantic Web*, 15(1):21–82.

Andreea Iana, Alexander Grote, Kaharina Ludwig, Mehwish Alam, Philipp Müller, Christof Weinhardt, Harald Sack, and Heiko Paulheim. 2022b. Geneg: German news knowledge graph.

Ilija Ilievski and Sujoy Roy. 2013. Personalized News Recommendation based on Implicit Feedback. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, pages 10–15.

Lise Jakobsen, Anna Johanne Holden, Önder Gürcan, and Özlem Özgöbek. 2025. Agent-Based Exploration of Recommendation Systems in Misinformation Propagation. *arXiv preprint arXiv:2507.21724*.

Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press.

Sepehr Janghorbani and Gerard De Melo. 2023. Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision–Language Models. *arXiv preprint arXiv:2303.12734*.

Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction*. Cambridge University Press.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Jacob Jensen, Suresh Naidu, Ethan Kaplan, Laurence Wilse-Samson, David Gergen, Michael Zuckerman, and Arthur Spirling. 2012. Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech [with Comments and Discussion]. *Brookings Papers on Economic Activity*, pages 1–81.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

Youchun Ji, Wenxing Hong, Yali Shangguan, Huan Wang, and Jing Ma. 2016. Regularized Singular Value Decomposition in News Recommendation System. In *2016 11th International Conference on Computer Science & Education (ICCSE)*, pages 621–626. IEEE.

Zhenyan Ji, Mengdan Wu, Hong Yang, and José Enrique Armendáriz Íñigo. 2021. Temporal Sensitive Heterogeneous Graph Neural Network for News Recommendation. *Future Generation Computer Systems*, 125:324–333.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM computing surveys*, 55(12):1–38.

Qinglin Jia, Jingjie Li, Qi Zhang, Xiuqiang He, and Jieming Zhu. 2021. RMBERT: News Recommendation via Recurrent Reasoning Memory Network over BERT. In *Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1773–1777.

Junxiang Jiang. 2023. TADI: Topic-aware Attention and Powerful Dual-encoder Interaction for Recall in News Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15647–15658.

Kevin Joseph and Hui Jiang. 2019. Content based News Recommendation via Shortest Entity Distance over Knowledge Graphs. In *Companion proceedings of the 2019 world wide web conference*, pages 690–699.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Michael Jugovac, Dietmar Jannach, and Mozhgan Karimi. 2018. Streamingrec: A framework for benchmarking stream-based news recommenders. In *Proceedings of the 12th ACM conference on recommender systems*, pages 269–273.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Lucie-Aimée Kaffee, Russa Biswas, C Maria Keet, Edlira Kalemi Vakaj, and Gerard de Melo. 2023. Multilingual Knowledge Graphs and Low-Resource Languages: A Review. *Transactions on Graph Data and Knowledge*, 1(1):10–1.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language Identification for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News Recommender Systems–Survey and Roads Ahead. *Information Processing & Management*, 54(6):1203–1227.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258.

Gabriella Kazai, Iskander Yusof, and Daoud Clarke. 2016. Personalised News and Blog Recommendations based on User Location, Facebook and Twitter User Profiling. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1129–1132.

Przemysław Kazienko and Erik Cambria. 2024. Towards Responsible Recommender Systems. *IEEE Intelligent Systems*, 39(3).

Tanzila Kehkashan, Raja Adil Riaz, Ahmad Sami Al-Shamayleh, Adnan Akhunzada, Noman Ali, Muhammad Hamza, and Faheem Akbar. 2025. AI-Generated Text Detection: A Comprehensive Review of Methods, Datasets, and Applications. *Computer Science Review*, 58:100793.

Krishnaram Kenthapadi, Benjamin Le, and Ganesh Venkataraman. 2017. Personalized Job Recommendation System at LinkedIn: Practical Challenges and Lessons Learned. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, page 346–347, New York, NY, USA. Association for Computing Machinery.

Hikmat Ullah Khan, Anam Naz, Fawaz Khaled Alarfaj, and Naif Almusallam. 2025. Identifying artificial intelligence-generated content using the distilbert transformer and nlp techniques. *Scientific Reports*, 15(1):20366.

Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018. Weave&Rec: A Word Embedding based 3-D Convolutional Network for News Recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Managemen*, pages 1855–1858.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 4110–4124.

Dang Kieu, Delvin Ce Zhang, Minh Duc Nguyen, Min Xu, Qiang Wu, and Dung D Le. 2025a. Enhancing News Recommendation with Hierarchical LLM Prompting. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 3072–3075.

Hai-Dang Kieu, Minh-Duc Nguyen, Thanh-Son Nguyen, and Dung D Le. 2025b. Keyword-driven Retrieval-Augmented Large Language Models for Cold-start User Recommendations. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2717–2721.

Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. 2013. The Plista Dataset. In *Proceedings of the 2013 international news recommender systems workshop and challenge*, pages 16–23.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Joel Kiskola, Henrik Rydenfelt, Thomas Olsson, Lauri Haapanen, Noora Vänttinen, Matti Nelimarkka, Minna Vigren, Salla-Maaria Laaksonen, and Tuukka Lehtiniemi. 2025. Generative AI and News Consumption: Design Fictions and Critical Analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Max Klabunde, Mehdi Ben Amor, Michael Granitzer, and Florian Lemmerich. 2023. Towards Measuring Representational Similarity of Large Language Models. In *UniReps: the First Workshop on Unifying Representations in Neural Models*.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. 2025. Similarity of Neural Network Models: A Survey of Functional and Representational Measures. *ACM Computing Surverys.*, 57(9).

Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A Survey on Popularity Bias in Recommender Systems. *User Modeling and User-Adapted Interaction*, 34(5):1777–1834.

Peter Knees, Julia Neidhardt, and Irina Nalis. 2024. Recommender Systems: Techniques, Effects, and Measures Toward Pluralism and Fairness. *Hannes Werthner· Carlo Ghezzi· Jeff Kramer· Julian Nida-Rümelin· Bashar Nuseibeh· Erich Prem·*, page 417.

Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-adapted Interaction*, 22(4):441–504.

Anton Korikov, Scott Sanner, Yashar Deldjoo, Zhankui He, Julian McAuley, Arnau Ramisa, Rene Vidal, Mahesh Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, et al. 2024. Large language model driven recommendation. *arXiv preprint arXiv:2408.10946*.

András Kornai. 2013. Digital Language Death. *PloS one*, 8(10):e77056.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Johannes Kruse, Kasper Lindskow, Michael Riis Andersen, Ryotaro Shimizu, Julian McAuley, Pierre-Alexandre Mattei, and Jes Frellsen. 2025. Normative Alignment of Recommender Systems via Internal Label Shift. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 1240–1245.

Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. EB-NeRD A Large-scale Dataset for News Recommendation. In *Proceedings of the Recommender Systems Challenge 2024*, pages 1–11.

Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Computing Surveys (CSUR)*, 53(1):12:1–12:37.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Maciej Kula. 2017. Spotlight. `https://github.com/maciejkula/spotlight`.

Ankit Kumar, Piyush Makhija, and Anuj Gupta. 2020. Noisy Text Data: Achilles' Heel of BERT. *arXiv preprint arXiv:2003.12932*.

Vaibhav Kumar, Dhruv Khattar, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Neural Architecture for News Recommendation. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*.

Per E Kummervold, Javier De La Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29.

Jessica Kunert and Neil Thurman. 2019. The Form of Content Personalisation at Mainstream, Transatlantic News Outlets: 2010–2016. *Journalism Practice*, 13(7):759–780.

Ken Lang. 1995. Newsweeder: Learning to Filter Netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier.

Sara Latifi and Dietmar Jannach. 2022. Streaming Session-Based Recommendation: When Graph Neural Networks meet the Neighborhood. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 420–426.

Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Konstantina Lazaridou and Ralf Krestel. 2016. Identifying Political Bias in News Articles. *Bulletin of the IEEE TCDL*, 12.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196. PMLR.

Olivia Le Poidevin. 2025. AI Assistants Make Widespread Errors About the News, New Research Shows. *Reuters*. [Online; accessed 16-January-2026].

Dongho Lee, Byungkook Oh, Seungmin Seo, and Kyong-Ho Lee. 2020a. News Recommendation with Topic-Enriched Knowledge Graphs. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 695–704.

Guang-He Lee and Yun-Nung Chen. 2017. MUSE: Modularizing Unsupervised Sense Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 327–337.

Hong Joo Lee and Sung Joo Park. 2007. MONERS: A News Recommender for the Mobile Web. *Expert Systems with Applications*, 32(1):143–150.

Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1073–1082.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020b. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4):1234–1240.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected? In *Proceedings of the 15th ACM conference on recommender systems*, pages 601–606.

Effi Levi, Guy Mor, Shaul Shenhav, and Tamir Sheafer. 2020. Compres: A dataset for narrative structure in news. *arXiv preprint arXiv:2007.04874*.

M Paul Lewis, Gary F Simons, and Charles D Fennig. 2009. Ethnologue: Languages of the World, Dallas, Texas: SIL International. *Online version: http://www. ethnologue. com*, 12(12):2010.

Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: Multi-Interest Matching Network for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 343–352.

Jingyuan Li, Yue Zhang, Xuan Lin, Xinxing Yang, Ge Zhou, Longfei Li, Hong Chen, and Jun Zhou. 2023a. TAML: Time-Aware Meta Learning for Cold-Start Problem in News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2415–2419.

Lei Li and Tao Li. 2013. News Recommendation via Hypergraph Learning: Encapsulation of User Behavior and News Content. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 305–314.

Lei Li, Ding-Ding Wang, Shun-Zhi Zhu, and Tao Li. 2011a. Personalized News Recommendation: A Review and an Experimental Investigation . *Journal of computer science and technology*, 26(5):754–766.

Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. 2011b. SCENE: A Scalable Two-stage Personalized News Recommendation System. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 125–134.

Lei Li, Li Zheng, and Tao Li. 2011c. Logo: A Long-short User Interest Integration in Personalized News Recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 317–320.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A Contextual-bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670.

Miaomiao Li and Licheng Wang. 2019. A Survey on Personalized News Recommendation Technology. *IEEE Access*, 7:145861–145879.

Runze Li, Lucien Heitz, Oana Inel, and Abraham Bernstein. 2025. D-RDW: Diversity-Driven Random Walks for News Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 558–563.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023b. A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, and Fake News. In *CEUR Workshop Proceedings*, volume 3561. CEUR-WS.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023c. Exploring Fine-tuning ChatGPT for News Recommendation. *arXiv preprint arXiv:2311.05850*.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023d. PBNR: Prompt-based News Recommender System. *arXiv preprint arXiv:2304.07862*.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Prompt-Based Generative News Recommendation (PGNR): Accuracy and Controllability. In *European Conference on Information Retrieval*, pages 66–79. Springer.

Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021. More Robust Dense Retrieval with Contrastive Dual Learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 287–296.

Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards Better Representation Learning for Personalized News Recommendation: a Multi-Channel Deep Fusion Approach. In *IJCAI*, pages 3805–3811.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1478–1484.

Chenjun Ling, Ben Steichen, and Silvia Figueira. 2020. Multilingual News–An Investigation of Consumption, Querying, and Search Result Selection Behaviors. *International Journal of Human–Computer Interaction*, 36(6):516–535.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On Cross-lingual Retrieval with Multilingual Text Encoders. *Information Retrieval Journal*, 25(2):149–183.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing Languages as Typological, Geographical, and Phylogenetic Vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Alexander Liu and Samuel Yang. 2022. Masked Autoencoders As The Unified Learners For Pre-Trained Sentence Representation. *arXiv preprint arXiv:2208.00231*.

Bing Liu et al. 2010a. Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Danyang Liu, Ting Bai, Jianxun Lian, Xin Zhao, Guangzhong Sun, Ji-Rong Wen, and Xing Xie. 2019a. News Graph: An Enhanced Knowledge Graph for News Recommendation. In *KaRS@ CIKM*, pages 1–7.

Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. 2020a. KRED: Knowledge-aware Document Representation for News Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 200–209.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021a. Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010b. Personalized News Recommendation Based on Click Behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, page 31–40, New York, NY, USA. Association for Computing Machinery.

Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A Shapiro, and Mustafa Bilgic. 2021b. The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. In *Proceedings of the Web Conference 2021*, pages 3791–3801.

Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024a. Multimodal Recommender Systems: A Survey. *ACM Computing Surveys*, 57(2):1–17.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024b. ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 452–461.

Qijiong Liu, Lu Fan, and Xiao-Ming Wu. 2025. Legommenders: A Comprehensive Content-Based Recommendation Library with LLM Support. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 753–756.

Ruixuan Liu, Yang Cao, Yanlin Wang, Lingjuan Lyu, Yun Chen, and Hong Chen. 2023. PrivateRec: Differentially Private Model Training and Online Serving for Federated News Recommendation. In *Proceedings of the 29th acm sigkdd conference on knowledge discovery and data mining*, pages 4539–4548.

Ruochen Liu, Hao Chen, Yuanchen Bei, Qijie Shen, Fangwei Zhong, Senzhang Wang, and Jianxin Wang. 2024c. Fine Tuning Out-of-Vocabulary Item Recommendation with User Sequence Imagination. *Advances in Neural Information Processing Systems*, 37:8930–8955.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021c. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

Zhan Liu, Shaban Shabani, Nicole Glassey Balet, and Maria Sokhn. 2019c. Detection of Satiric News on Social Media: Analysis of the Phenomenon with a French Dataset. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6. IEEE.

Andreas Lommatzsch, Benjamin Kille, and Sahin Albayrak. 2017. Incorporating Context and Trends in News Recommender Systems. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 1062–1068, New York, NY, USA. Association for Computing Machinery.

Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, et al. 2025. Bridging the Data Provenance Gap Across Text, Speech, and Video. In *The Thirteenth International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond Optimizing for Clicks: Incorporating Editorial Values in News Recommendation. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 145–153.

Sebastian Lubos, Alexander Felfernig, and Markus Tautschnig. 2023. An Overview of Video Recommender Systems: State-of-the-art and Research Issues. *Frontiers in big Data*, 6:1281614.

Joel Pinho Lucas, João Felipe Guedes da Silva, and Letícia Freire Figueiredo. 2023. NPR: a News Portal Recommendations Dataset. In *Proceedings of the The First Workshop on the Normative Design and Evaluation of Recommender Systems (NORMalize 2023), co-located with the ACM Conference on Recommender Systems 2023 (RecSys 2023)*.

Cornelius A Ludmann. 2017. Recommending News Articles in the CLEF News Recommendation Evaluation Lab with the Data Stream Management System Odysseus. In *CLEF (Working Notes)*.

Katharina Ludwig, Alexander Grote, Andreea Iana, Mehwish Alam, Heiko Paulheim, Harald Sack, Christof Weinhardt, and Philipp Müller. 2023. Divided by the Algorithm? The (Limited) Effects of Content- and Sentiment-Based News Recommendation on Affective, Ideological, and Perceived Polarization. *Social Science Computer Review*, 41(6):2188–2210.

Katharina Ludwig and Philipp Müller. 2022. Does Social Media Use Promote Political Mass Polarization? A Structured Literature Review. *Questions of Communicative Change and Continuity. In Memory of Wolfram Peiser*, pages 118–166.

Hangzai Luo, Jianping Fan, and Daniel A Keim. 2008. Personalized News Video Recommendation. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 1001–1002.

Yunze Luo, Yuezihan Jiang, Yinjie Jiang, Gaode Chen, Jingchi Wang, Kaigui Bian, Peiyi Li, and Qi Zhang. 2025. Online Item Cold-Start Recommendation with Popularity-Aware Meta-Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 927–937.

Emma Lurie and Eni Mustafaraj. 2019. Opening Up the Black Box: Auditing Google's Top Stories Algorithm. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*, volume 32, pages 376–381.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024a. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612.

Yougang Lyu, Xiaoyu Zhang, Zhaochun Ren, and Maarten de Rijke. 2024b. Cognitive Biases in Large Language Models for News Recommendation. In *Proceedings of the 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@RecSys 2024), October 2024, Bari, Italy.*

Mingyuan Ma, Sen Na, Hongyu Wang, Congzhou Chen, and Jin Xu. 2022. The graph-based behavior-aware recommendation for interactive news. *Applied Intelligence*, 52(2):1913–1929.

Ian MacKenzie, Chris Meyer, and Steve Noble. 2013. How retailers can keep up with consumers. *McKinsey&Company*. [Online; accessed 12-August-2025].

Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. 2023. On Popularity Bias of Multimodal-aware Recommender Systems: A Modalities-driven Analysis. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval*, pages 59–68.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2145–2148, New York, NY, USA. Association for Computing Machinery.

Zhiming Mao, Xingshan Zeng, and Kam-Fai Wong. 2021. Neural news recommendation with collaborative news encoding and structural user encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 46–55.

Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. 2021. Stance Detection in German News Articles. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 66–77, Dominican Republic. Association for Computational Linguistics.

Philip May. 2020. Cross English & German RoBERTa for Sentence Embeddings.

Jennifer McCoy, Tahmina Rahman, and Murat Somer. 2018. Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities. *American Behavioral Scientist*, 62(1):16–42.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Denis McQuail. 1992. *Media Performance: Mass Communication and the Public Interest*, volume 144. Sage London.

Alessandro B Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating Gender Fairness of Recommendation Algorithms in the Music Domain. *Information Processing & Management*, 58(5):102666.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *Advances in neural information processing systems*, 35:17359–17372.

Lingkang Meng, Chongyang Shi, Shufeng Hao, and Xiangrui Su. 2021. DCAN: Deep Co-Attention Network by Modeling User Preference and News Lifecycle for News Recommendation. In *International Conference on Database Systems for Advanced Applications*, pages 100–114. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender Systems and Their Ethical Challenges. *Ai & Society*, 35(4):957–967.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Amy Mitchell, Mark Jurkowitz, J Baxter Oliphant, and Elisa Shearer. 2020. Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable. *Pew Research Center*, 30.

Marnix Moerland, Frederik Hogenboom, Michel Capelle, and Flavius Frasincar. 2013. Semantics-based News Recommendation with SF-IDF+. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–8.

Lucas Möller and Sebastian Padó. 2022. Understanding the Relation of User and News Representations in Content-Based Neural News Recommendation. *Joint Proceedings of 10th International Workshop on News Recommendation and Analytics (INRA'22) and the Third International Workshop on Investigating Learning During Web Search (IWILDS'22) co-located with the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22)*.

Lucas Möller and Sebastian Padó. 2024. Explaining Neural News Recommendation with Attributions onto Reading Histories. *ACM Transactions on Intelligent Systems and Technology*.

Lynge Asbjørn Møller. 2022. Recommended for You: How Newspapers Normalise Algorithmic News Recommendation to Fit Their Gatekeeping Role. *Journalism Studies*, 23(7):800–817.

Armin Moradi, Nicola Neophytou, Florian Carichon, and Golnoosh Farnadi. 2025. Embedding Cultural Diversity in Prototype-based Recommender Systems. In *European Conference on Information Retrieval*, pages 17–31. Springer.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977.

Arvind Narayanan. 2023. Understanding social media recommendation algorithms: Towards a better informed debate on the effects of social media. Essay on Knight First Amendment Institute. [Online; accessed 04-November-2025].

Sarah Naseer, Christopher St. Aubin, and Michael Lipka. 2024a. English- and Spanish-Language News Consumption Among Hispanics. [Online; accessed: 15-September-2025].

Sarah Naseer, Christopher St. Aubin, and Michael Lipka. 2024b. How hispanic americans get their news. [Online; accessed 25-August-2025].

National Telecommunications and Information Administration. 2001. Falling through the Net: Defining the Digital Divide. In *The Digital Divide: Facing a Crisis or Creating a Myth?* The MIT Press.

Fatemeh Nazary, Ali Tourani, Yashar Deldjoo, and Tommaso Di Noia. 2025. ViLLA-MMBench: A Unified Benchmark Suite for LLM-Augmented Multimodal Movie Recommendation. *arXiv preprint arXiv:2508.04206*.

Nic Newman and Federica Cherubini. 2025. Journalism, Media, and Technology Trends and Predictions. *Oxford: Reuters Institute for the Study of Journalism*.

Nic Newman, Richard Fletcher, Craig T. Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. 2024. *Digital News Report 2024*. Reuters Institute for the Study of Journalism.

Nic Newman, A Ross Arguedas, Craig T Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. *Digital News Report 2025*. Reuters Institute for the Study of Journalism.

Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 677–686, New York, NY, USA. Association for Computing Machinery.

Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan & Claypool Publishers.

Dan Saattrup Nielsen. 2023. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In *The 24rd Nordic Conference on Computational Linguistics*.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's Paradox of 'Low-Resource' Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Yunseok Noh, Yong-Hwan Oh, and Seong-Bae Park. 2014. A Location-based Personalized News Recommendation. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 99–104. IEEE.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based News Recommendation for Millions of Users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1933–1942.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.

Andreas L. Opdahl, Tareq Al-Moslmi, Duc-Tien Dang-Nguyen, Marc Gallofré Ocaña, Bjørnar Tessem, and Csaba Veres. 2022. Semantic Knowledge Graphs for the News: A Review. *ACM Computing Surveys*.

OpenAI. 2022. Introducing chatgpt. [Online; accessed 13-August-2025].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto,

Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yoshihiko Ozaki, Yuki Tanigaki, Shuhei Watanabe, and Masaki Onishi. 2020. Multiobjective Tree-structured Parzen Estimator for Computationally Expensive Optimization Problems. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 533–541.

Özlem Özgöbek, Jon Atle Gulla, and R Cenk Erdur. 2014. A Survey on Challenged and Methods in News Recommendation. In *International Conference on Web Information Systems and Technologies*, volume 2, pages 278–285. SCITEPRESS.

Makbule Gulcin Ozsoy. 2024. Multilingual Prompts in LLM-Based Recommenders: Performance Across Languages. *arXiv preprint arXiv:2409.07604*.

Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. penguin UK.

Ali Hakimi Parizi and Mohammad Kazemifard. 2015. Emotional News Recommender System. In *2015 Sixth International Conference of Cognitive Science (ICCS)*, pages 37–41. IEEE.

Ali Hakimi Parizi, Mohammad Kazemifard, and Mohsen Asghari. 2016. EmoNews: An Emotional News Recommender System. *Journal of Digital Information Management*, 14(6).

Choongwon Park. 2025. LLM as User Simulator: Towards Training News Recommender without Real User Interactions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3080–3084.

Hyunji Hayley Park, Katherine J Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology Matters: A Multilingual Language Modeling Analysis . *Transactions of the Association for Computational Linguistics*, 9:261–276.

Keunchan Park, Jisoo Lee, and Jaeho Choi. 2017a. Deep Neural Networks for News Recommendations. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2255–2258.

Seyeon Park and Xiaoli Nan. 2025. Generative ai and misinformation: A scoping review of the role of generative ai in the generation, detection, mitigation, and impact of misinformation. *AI & SOCIETY*, pages 1–15.

Yoojin Park, Jinoh Oh, and Hwanjo Yu. 2017b. RecTime: Real-time Recommender System for Online Broadcasting. *Information Sciences*, 409:1–16.

Anish Patankar, Joy Bose, and Harshit Khanna. 2019. A Bias Aware News Recommendation System. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 232–238. IEEE.

Patricia Ramos. 2025. New data on detailed languages spoken at home and the ability to speak english. [Online; accessed 25-August-2025].

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337.

Andreas Peintner. 2023. Sequential Recommendation Models: A Graph-based Perspective. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1295–1299.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Owen Phelan, Kevin McCarthy, and Barry Smyth. 2009. Using Twitter to Recommend Real-time Topical News. In *Proceedings of the third ACM conference on Recommender systems*, pages 385–388.

Alessandro Piscopo, Anna McGovern, Lianne Kerlin, North Kuras, James Fletcher, Calum Wiggins, and Megan Stamper. 2024. Recommenders With Values: Developing Recommendation Engines in a Public Service Organization. *Knight First Amendment Institute at Columbia University: Essays and Scholarship*. Accessed: 12.08.2025.

Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in Rankings and Recommendations: An Overview. *The VLDB Journal*, pages 1–28.

Gustavo P Polleti, Douglas L de Souza, and Fabio G Cozman. 2025. Generating Explanations for Knowledge-Aware Conversational Recommendation Systems. *User Modeling and User-Adapted Interaction*, 35(3):12.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *International Conference on Learning Representations*.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021a. Personalized News Recommendation with Knowledge-aware Interactive Matching. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–70.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021b. Personalized News Recommendation with Knowledge-aware Interactive Matching. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–70.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021c. PP-Rec: News Recommendation with Personalized User Interest and Time-aware News Popularity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5457–5467.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022a. FUM: Fine-grained and Fast User Modeling for News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1974–1978.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022b. News Recommendation with Candidate-aware User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1917–1921.

Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2020. Privacy-Preserving News Recommendation Model Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1423–1432.

Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2021d. Uni-FedRec: A Unified Privacy-Preserving News Recommendation Framework for Model Training and Online Serving. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1438–1448.

Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021e. HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5446–5456.

Yongye Qian, Pengpeng Zhao, Zhixu Li, Junhua Fang, Lei Zhao, Victor S Sheng, and Zhiming Cui. 2019. Interaction Graph Neural Network for News Recommendation. In *International Conference on Web Information Systems Engineering*, pages 599–614. Springer.

Jun Qin, Sitong Liu, and Junjie Liang. 2025. LLM-Based News Recommendation System With Multi-Granularity News Content Fusion and Dual-View User Interest Perception. *IEEE Access*.

Junyang Rao, Aixia Jia, Yansong Feng, and Dongyan Zhao. 2013. Taxonomy based Personalized News Recommendation: Novelty and Diversity. In *Web Information Systems Engineering–WISE 2013: 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part I 14*, pages 209–218. Springer.

Muhammad Arslan Rauf, Mian Muhammad Yasir Khalil, Muhammad Ahmad Nawaz Ul Ghani, Weidong Wang, Qingxian Wang, and Junaid Hassan. 2024a. ZS-CEBE: Leveraging Zero-shot Cross and Bi-encoder Architecture for Cold-start News Recommendation. *Signal, Image and Video Processing*, 18(8):6455–6467.

Muhammad Arslan Rauf, Mian Muhammad Yasir Khalil, Weidong Wang, Qingxian Wang, Muhammad Ahmad Nawaz Ul Ghani, and Junaid Hassan. 2024b. BCE4ZSR: Bi-encoder Empowered by Teacher Cross-encoder for Zero-shot Cold-start News Recommendation. *Information Processing & Management*, 61(3):103686.

Kumar Ravi and Vadlamani Ravi. 2015. A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications. *Knowledge-based systems*, 89:14–46.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573.

Shaina Raza. 2023. Bias Reduction News Recommendation System. *Digital*, 4(1):92–103.

Shaina Raza and Chen Ding. 2020. A Regularized Model to Trade-off between Accuracy and Diversity in a News Recommender System. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 551–560. IEEE.

Shaina Raza and Chen Ding. 2021. Deep Dynamic Neural Network to trade-off between Accuracy and Diversity in a News Recommender System. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5246–5256.

Shaina Raza and Chen Ding. 2022. News Recommender System: A Review of Recent Progress, Challenges, and Opportunities. *Artificial Intelligence Review*, pages 1–52.

Elizabeth Reid. 2023. Supercharging Search with generative AI. Blog post on Google Products website. [Online; accessed 04-November-2025].

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020a. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Nils Reimers and Iryna Gurevych. 2020b. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461.

Steffen Rendle, Li Zhang, and Yehuda Koren. 2019. On the Difficulty of Evaluating Baselines: A Study on Recommender Systems. *arXiv preprint arXiv:1905.01395*.

Paul Resnick, R Kelly Garrett, Travis Kriplean, Sean A Munson, and Natalie Jomini Stroud. 2013. Bursting your (filter) bubble: strategies for promoting diverse exposure. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, pages 95–100.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–186.

Paul Resnick and Hal R Varian. 1997. Recommender Systems. *Communications of the ACM*, 40(3):56–58.

Myrthe Reuver, Antske Fokkens, Suzan Verberne, H Toivonen, and M Boggia. 2021. No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems. *Proceedings of the EACL Hackashop on news media content analysis and automated report generation*, pages 45–55.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2022. *Recommender Systems Handbook*. Springer New York, NY.

Khumukcham Robindro, Kshetrimayum Nilakanta, Deepen Naorem, and Ningthoujam Gourakishwar Singh. 2017. An Unsupervised Content Based News Personalization using Geolocation Information. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 128–132. IEEE.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building Event-Centric Knowledge Graphs from News. *Journal of Web Semantics*.

Qin Ruan, Brian Mac Namee, and Ruihai Dong. 2023. Unveiling the Relationship Between News Recommendation Algorithms and Media Bias: A Simulation-Based Analysis of the Evolution of Bias Prevalence. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 210–215. Springer.

Qin Ruan, Brian Mac Namee, and Ruihai Dong. 2024a. The Effects of Media Bias on News Recommendations. *IEEE Access*, 12:83391–83404.

Qin Ruan, Jin Xu, Susan Leavy, Brian Mac Namee, and Ruihai Dong. 2024b. Rewriting Bias: Mitigating Media Bias in News Recommender Systems through Automated Rewriting. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 67–77.

Xie Runfeng, Cui Xiangyang, Yan Zhou, Wang Xin, Xuan Zhanwei, Zhang Kai, et al. 2023. LKPNR: LLM and KG for Personalized News Recommendation Framework. *arXiv preprint arXiv:2308.12028*.

Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H Schoenmacker. 2022. Closing the Gender Wage Gap: Adversarial Fairness in Job Recommendation. In *RecSys in HR'22: The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems, September 18–23, 2022, Seattle, USA*. CEUR-WS. org.

Jenna Russell, Marzena Karpinska, Destiny Akinode, Katherine Thai, Bradley Emi, Max Spero, and Mohit Iyyer. 2025. Ai use in american newspapers is widespread, uneven, and rarely disclosed. *arXiv preprint arXiv:2510.18774*.

Seongeun Ryu, Yunyong Ko, and Sang-Wook Kim. 2025. Is This News Still Interesting to You?: Lifetime-aware Interest Matching for News Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2515–2524.

Alan Said. 2024. Recommender Systems for Social Good: The Role of Accountability and Sustainability. In *International Workshop on Recommender Systems for Sustainability and Social Good*, pages 1–4. Springer.

Alan Said and Alejandro Bellogín. 2014. Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 129–136.

Aghiles Salah, Quoc-Tuan Truong, and Hady W Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *The Journal of Machine Learning Research*, 21(1):3803–3807.

William Samuelson and Richard Zeckhauser. 1988. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1(1):7–59.

Pablo Sánchez, Alejandro Bellogín, and Ludovico Boratto. 2023. Bias Characterization, Assessment, and Mitigation in Location-based Recommender Systems. *Data Mining and Knowledge Discovery*, 37(5):1885–1929.

Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. *arXiv preprint cs/0306050*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large Language Models are Competitive Near Cold-start Recommenders for Language- and Item-based Preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 890–896.

T.Y.S.S Santosh, Avirup Saha, and Niloy Ganguly. 2020. MVL: Multi-view Learning for News Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1873–1876.

Fabian David Schmidt, Markus Dietsche, Simone Paolo Ponzetto, and Goran Glavaš. 2019. SEAGLE: A Platform for Comparative Evaluation of Semantic Encoders for Information Retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 199–204.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't Stop Fine-Tuning: On Training Regimes for Few-Shot Cross-Lingual Transfer with Multilingual Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*, volume 39. Cambridge University Press Cambridge.

Khadiga Seddik. 2024. Exploring the Ethical Challenges of AI and Recommender Systems in the Democratic Public Sphere. In *Norsk IKT-konferanse for forskning og utdanning*, 1.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.

Mete Sertkan and Julia Neidhardt. 2022a. Diversifying Sentiments in News Recommendation. In *Proceedings of the 2nd Perspectives on the Evaluation of Recommender Systems Workshop*.

Mete Sertkan and Julia Neidhardt. 2022b. Exploring Expressed Emotions for Neural News Recommendation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 22–28.

Mete Sertkan and Julia Neidhardt. 2023. On the Effect of Incorporating Expressed Emotions in News Articles on Diversity within Recommendation Models. *decision-making*, 3:11.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – A Multilingual Cross-domain Fact Check News Dataset for COVID-19 . In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, and Yong Li. 2024. Improving Item-side Fairness of Multimodal Recommendation via Modality Debiasing. In *Proceedings of the ACM Web Conference 2024*, pages 4697–4705.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.

Heng-Shiou Sheu and Sheng Li. 2020. Context-aware Graph Embedding for Session-based News Recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 657–662.

Hao Shi, Zi-Jiao Wang, and Lan-Ru Zhai. 2022. DCAN: Diversified News Recommendation with Coverage-Attentive Networks. *arXiv preprint arXiv:2206.02627*.

Shaoyun Shi, Weizhi Ma, Zhen Wang, Min Zhang, Kun Fang, Jingfang Xu, Yiqun Liu, and Shaoping Ma. 2021. WG4Rec: Modeling Textual Content with Word Graph for News Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Managemen*, pages 1651–1660.

Karthik Shivaram, Ping Liu, Matthew Shapiro, Mustafa Bilgic, and Aron Culotta. 2022. Reducing Cross-topic Political Homogenization in Content-based News Recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 220–228.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big data*, 8(3):171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. 2024. LoRA vs Full Fine-tuning: An Illusion of Equivalence. *arXiv preprint arXiv:2410.21228*.

Felix M Simon. 2024. Artificial Intelligence in the News: How AI Retools, Rationalizes, and Reshapes Journalism and the Public Arena. Technical report, Tow Center for Digital Journalism, Columbia University.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing Digital Language Support on a Global Scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Brent Smith and Greg Linden. 2017. Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing*, 21(3):12–18.

Joan E. Solsman. 2018. YouTube's AI is the puppet master over most of what you watch. CNET. [Online; accessed 12-August-2025].

Jeong-Woo Son, A-Yeong Kim, and Seong-Bae Park. 2013. A Location-based News Article Recommendation with Explicit Localized Semantic Analysis. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 293–302.

Zhanghan Song, Dian Zhang, Xiaochuan Shi, Wei Li, Chao Ma, and Libing Wu. 2021. DEN-DQL: Quick Convergent Deep Q-Learning with Double Exploration Networks for News Recommendation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Karen Sparck Jones. 1972. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of documentation*, 28(1):11–21.

Timo Spinde, Felix Hamborg, and Bela Gipp. 2020. An Integrated Approach to Detect Media Bias in German News Articles. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 505–506.

Ben Steichen, M Rami Ghorab, Alexander O'Connor, Séamus Lawless, and Vincent Wade. 2014. Towards Personalized Multilingual Information Access - Exploring the Browsing and Search Behavior of Multilingual Users. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 435–446. Springer, Cham.

Catherine Stinson. 2022. Algorithms are not neutral: Bias in collaborative filtering. *AI and Ethics*, 2(4):763–770.

Natalie Jomini Stroud. 2017. Selective Exposure Theories. In *The Oxford handbook of political communication*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

Li Sun, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio, and Cha Zhang. 2023. From Characters to Words: Hierarchical Pre-trained Language Model for Open-vocabulary Language Understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3605–3620.

Yumin Sun, Fangzhou Yi, Cheng Zeng, Bing Li, Peng He, Jinxia Qiao, and Yinghui Zhou. 2021. A Hybrid Approach to News Recommendation Based on Knowledge Graph and Long Short-Term User Preferences. In *2021 IEEE International Conference on Services Computing (SCC)*, pages 165–173. IEEE.

Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. 2020. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 23–32.

Cass R Sunstein. 2001. *Echo Chambers: Bush v. Gore, Impeachment, and Beyond*. Princeton University Press Princeton, NJ.

Mozhgan Tavakolifard, Jon Atle Gulla, Kevin C Almeroth, Jon Espen Ingvaldesn, Gaute Nygreen, and Erik Berg. 2013. Tailored News in the Palm of your Hand: A Multi-perspective Transparent Approach to News Recommendation. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 305–308.

Wilson L Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism quarterly*, 30(4):415–433.

Bing Index Team. 2013. Understand Your World with Bing. *Microsoft Bing Blogs*. [Online; accessed 11-August-2025].

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex

Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. 2025. FreshStack: Building Realistic Benchmarks for Evaluating Retrieval on Technical Documents. *arXiv preprint arXiv:2504.13128*.

Yu Tian, Yuhao Yang, Xudong Ren, Pengfei Wang, Fangzhao Wu, Qian Wang, and Chenliang Li. 2021. Joint Knowledge Pruning and Recurrent Graph Convolution for News Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 51–60.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

TikTok. 2020. How TikTok recommends videos #ForYou. [Online; accessed 04-November-2025].

Nava Tintarev, Bart P Knijnenburg, and Martijn C Willemsen. 2024. Measuring the Benefit of Increased Transparency and Control in News Recommendation. *Ai Magazine*, 45(2):212–226.

Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. 2018. Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. In *Adjunct publication of the 26th conference on user modeling, adaptation and personalization*, pages 7–13.

Yi-Qi Tong, Qian-Qi Liu, Wei Guo, Hong-Rui Niu, Fu-Zhen Zhuang, De-Qing Wang, and Jun Gao. 2025. PCRec: A Multi-Interest News Recommendation Framework with Prompt-Guided Cross-View Contrastive Learning. *Journal of Computer Science and Technology*, 40(4):1079–1093.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *Advances in Neural Information Processing Systems*, 33:2207–2219.

Dai Hoang Tran, Salma Hamad, Munazza Zaib, Abdulwahab Aljubairy, Quan Z Sheng, Wei Emma Zhang, Nguyen H Tran, and Nguyen Lu Dang Khoa. 2021. Deep News Recommendation with Contextual User Profiling and Multifaceted Article Representation. In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part II 22*, pages 237–251. Springer.

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *science*, 211(4481):453–458.

Twitter. 2023. Twitter's Recommendation Algorithm. [Online; accessed 12-August-2025].

United Nations. 1948. Universal declaration of human rights. [Online; accessed 18-August-2025].

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A Multilingual Multi-Target Dataset for Stance Detection. *CoRR*, abs/2003.08385.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of machine learning research*, 9(11).

Jan Van Dijk. 2020. *The Digital Divide*. Polity Press, Cambridge.

Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and Use of the Simple Event Model (SEM). *Journal of Web Semantics*, 9(2):128–136.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

Paul S Voakes, Jack Kapfer, David Kurpius, and David Shano-yeon Chern. 1996. Diversity in the News: A Conceptual and Methodological Framework. *Journalism & Mass Communication Quarterly*, 73(3):582–593.

Ellen M Voorhees et al. 1999. The TREC-8 Question Answering Track Report. In *Trec*, volume 99, pages 77–82.

Vladislav Vorotilov and Ilnur Shugaepov. 2023. Scaling the Instagram Explore recommendations system. Blog post on Engineering at Meta. [Online; accessed 04-November-2025].

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 2016. NewsReader: Using Knowledge Resources in a Cross-lingual Reading Machine to Generate More Knowledge from Massive streams of News. *Special Issue Knowledge-Based Systems, Elsevier*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85.

Sanne Vrijenhoek. 2023. Do you MIND? Reflections on the MIND Dataset for Research on Diversity in News Recommendations. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 147–154. Springer.

Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 208–219.

Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 173–183.

W3Techs. 2025. Usage Statistics of Content Languages for Websites. [Online; accessed 25-August-2025].

Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained Interest Matching for Neural News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 836–845.

Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018a. RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 417–426.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018b. DKN: Deep Knowledge-Aware Network for News Recommendation. In *Proceedings of the 2018 world wide web conference*, pages 1835–1844.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021a. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022a. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Lei Wang and Ee-Peng Lim. 2023. Zero-Shot Next-Item Recommendation using Large Pretrained Language Models. *arXiv preprint arXiv:2304.03153*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving Text Embeddings with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

Rongyao Wang, Veronica Liesaputra, and Zhiyi Huang. 2025. A Survey on LLM-based News Recommender Systems. *arXiv preprint arXiv:2502.09797*.

Rongyao Wang, Shoujin Wang, Wenpeng Lu, and Xueping Peng. 2022c. News Recommendation Via Multi-Interest News Sequence Modelling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7942–7946. IEEE.

Shuai Wang, Kun Zhang, Le Wu, Haiping Ma, Richang Hong, and Meng Wang. 2021b. Privileged Graph Distillation for Cold Start Recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1187–1196.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024c. RecMind: Large Language Model Powered Agent For Recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4351–4364.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022d. English Contrastive Learning Can Learn Universal Cross-lingual Sentence Embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3).

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Michael L Waskom. 2021. Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60):3021.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. PolyLM: An Open Source Polyglot Large Language Model. *arXiv preprint arXiv:2307.06018*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On Learning Universal Representations Across Languages. In *International Conference on Learning Representations*.

Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-start Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5382–5390.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012.

Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta. 2021. Recommender Systems and the Amplification of Extremist Content. *Internet Policy Review*, 10(2).

Wikimedia Commons. 2020. File:PercentWikipediasGraph.png. [Online; accessed 25-August-2025].

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz

Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020a. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural News Recommendation with Attentive Multi-View Learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3863–3869.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. NPA: Neural News Recommendation with Personalized Attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2576–2584.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019c. Neural News Recommendation with Topic-Aware News Representation. In *Proceedings of the 57th Annual meeting of the association for computational linguistics*, pages 1154–1159.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Tao Qi, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019d. Neural News Recommendation with Heterogeneous User Behavior. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4874–4883.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019e. Neural News Recommendation with Multi-Head Self-Attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6389–6394.

Chuhan Wu, Fangzhao Wu, Xiangnan He, and Yongfeng Huang. 2022a. DebiasGAN: Eliminating Position Bias in News Recommendation with Adversarial Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2933–2938.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2022b. Rethinking InfoNCE: How Many Negative Samples Do You Need? In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence ((IJCAI-22))*, pages 2509–2515.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2020b. Neural News Recommendation with Negative Feedback. *CCF Transactions on Pervasive Computing and Interaction*, 2:178–188.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2021a. User-as-Graph: User Modeling with Heterogeneous Graph Pooling for News Recommendation. In *IJCAI*, pages 1624–1630.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized News Recommendation: Methods and Challenges. *ACM Transactions on Information Systems*, 41(1):1–50.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020c. SentiRec: Sentiment Diversity-aware Neural News Recommendation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 44–53.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020d. User Modeling with Click Preference and Reading Satisfaction for News Recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3023–3029.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021b. Empowering News Recommendation with Pre-trained Language Models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1652–1656.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021c. MM-Rec: Multimodal News Recommendation. *arXiv preprint arXiv:2104.07407*.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022c. End-to-end Learnable Diversity-aware News Recommendation. *arXiv preprint arXiv:2204.00539*.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022d. Two Birds with One Stone: Unified Model Learning for Both Recall and Ranking in News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3474–3480.

Chuhan Wu, Fangzhao Wu, Tao Qi, Chenliang Li, and Yongfeng Huang. 2022e. Is News Recommendation a Sequential Recommendation Task? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2382–2386.

Chuhan Wu, Fangzhao Wu, Tao Qi, Chenliang Li, and Yongfeng Huang. 2022f. Is News Recommendation a Sequential Recommendation Task? In *Proceedings of the 45th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2382–2386.

Chuhan Wu, Fangzhao Wu, Tao Qi, Qi Liu, Xuan Tian, Jie Li, Wei He, Yongfeng Huang, and Xing Xie. 2022g. FeedRec: News Feed Recommendation with Various User Feedbacks. In *Proceedings of the ACM Web Conference 2022*, pages 2088–2097.

Chuhan Wu, Fangzhao Wu, Tao Qi, Wei-Qiang Zhang, Xing Xie, and Yongfeng Huang. 2022h. Removing AI's Sentiment Manipulation of Personalized News Delivery. *Humanities and Social Sciences Communications*, 9(1):1–9.

Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021d. Fairness-aware News Recommendation with Decomposed Adversarial Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4462–4469.

Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu. 2021e. NewsBERT: Distilling Pre-trained Language Model for Intelligent News Application. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3285–3295.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020e. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020f. Similarity Analysis of Contextual Word Representation Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4638–4655.

Junda Wu, Cheng-Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian McAuley. 2024a. CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3391–3401.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024b. A Survey on Large Language Models for Recommendation. *World Wide Web*, 27(5):60.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2022i. Graph Neural Networks in Recommender Systems: A Survey. *ACM Computing Surveys*, 55(5):1–37.

Xuansheng Wu, Huachi Zhou, Yucheng Shi, Wenlin Yao, Xiao Huang, and Ninghao Liu. 2024c. Could Small Language Models Serve as Recommenders? Towards Data-centric Cold-start Recommendation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3566–3575, New York, NY, USA. Association for Computing Machinery.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Youlin Wu, Yuanyuan Sun, Xiaokun Zhang, Haoxi Zhan, Bo Xu, Liang Yang, and Hongfei Lin. 2025. IP2: Entity-Guided Interest Probing for Personalized News Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 187–196.

Shitao Xiao, Zheng Liu, Yingxia Shao, Tao Di, Bhuvan Middha, Fangzhao Wu, and Xing Xie. 2022. Training Large-Scale News Recommenders with Pretrained Language Models in the Loop. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4215–4225.

Yingyuan Xiao, Pengqiang Ai, Ching-hsien Hsu, Hongya Wang, and Xu Jiao. 2015. Time-Ordered Collaborative Filtering for News Recommendation. *China Communications*, 12(12):53–62.

Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive Learning for Sequential Recommendation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1259–1273. IEEE.

Yanwei Xie, Weizhi Nie, Lanjun Wang, Hongshuo Tian, Changtai Shi, and An-An Liu. 2025. When Headlines Meet Minds: Empowering News Recommendations with Social Simulator. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6441–6450.

Hongyan Xu, Qiyao Peng, Hongtao Liu, Yueheng Sun, and Wenjun Wang. 2023. Group-based Personalized News Recommendation with Long-and Short-term Fine-Grained Matching. *ACM Transactions on Information Systems*.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. 2025. A Survey on Multimodal Recommender Systems: Recent Advances and Future Directions. *arXiv preprint arXiv:2502.15711*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiuqiang He, Xiaofei He, Tat-Seng Chua, and Fei Wu. 2021. Why Do We Click: Visual Impression-aware News Recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3881–3890.

Yuki Yada and Hayato Yamana. 2024. News Recommendation with Category Description by a Large Language Model. *arXiv preprint arXiv:2405.13007*.

Omry Yadan. 2019. Hydra - A Framework for Elegantly Configuring Complex Applications. Github.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Cheng-Zen Yang, Xiang Chen, and Ping-Jung Wu. 2008. Cross-Lingual News Group Recommendation Using Cluster-Based Cross-Training. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 13, Number 1, March 2008: Special Issue on Cross-Lingual Information Retrieval and Question Answering*, pages 41–60.

Chun Yang, Jianxiao Zou, JianHua Wu, Hongbing Xu, and Shicai Fan. 2022. Supervised Contrastive Learning for Recommendation. *Knowledge-Based Systems*, 258:109973.

JungAe Yang. 2016. Effects of Popularity-Based News Recommendations ("Most-Viewed") on Users' Exposure to Online News. *Media Psychology*, 19(2):243–271.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving Multilingual Sentence Embedding using Bi-directional Dual Encoder with Additive Margin Softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5370–5378.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2020a. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics: system demonstrations*, pages 87–94.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020b. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in neural information processing systems*, 32.

Kam Fung Yeung and Yanyan Yang. 2010. A Proactive Personalized Mobile News Recommendation System. In *2010 Developments in E-systems Engineering*, pages 207–212. IEEE.

Jingwei Yi, Fangzhao Wu, Chuhan Wu, Qifei Li, Guangzhong Sun, and Xing Xie. 2021a. Debiased-dRec: Bias-aware User Modeling and Click Prediction for Personalized News Recommendation. *arXiv preprint arXiv:2104.07360*.

Jingwei Yi, Fangzhao Wu, Chuhan Wu, Ruixuan Liu, Guangzhong Sun, and Xing Xie. 2021b. Efficient-FedRec: Efficient Federated Learning Framework for Privacy-Preserving News Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2824.

Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond Clicks: Dwell Time for Personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 113–120.

Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang, Zhaojun Ding, Yizhe Yang, Xuanjing Huang, and Shuicheng Yan. 2024. Automating Dataset Updates Towards Reliable and Timely Evaluation of Large Language Models. *Advances in Neural Information Processing Systems*, 37:17106–17132.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 974–983.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-Resource Languages Jailbreak GPT-4. In *NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR)*.

Puxuan Yu and James Allan. 2020. A Study of Neural Matching Models for Cross-lingual IR. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1637–1640.

Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi, and Qi Liu. 2022. Tiny-NewsRec: Effective and Efficient PLM-based News Recommendation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5478–5489.

Tianzi Zang, Yanmin Zhu, Haobing Liu, Ruohan Zhang, and Jiadi Yu. 2022. A Survey on Cross-domain Recommendation: Taxonomies, Methods, and Future Directions. *ACM Transactions on Information Systems*, 41(2):1–39.

Chiyu Zhang, Yifei Sun, Jun Chen, Jie Lei, Muhammad Abdul-Mageed, Sinong Wang, Rong Jin, Sem Park, Ning Yao, and Bo Long. 2024a. SPAR: Personalized Content-Based Recommendation via Long Engagement Attention. *arXiv preprint arXiv:2402.10555*.

Chiyu Zhang, Yifei Sun, Minghao Wu, Jun Chen, Jie Lei, Muhammad Abdul-Mageed, Rong Jin, Angli Liu, Ji Zhu, Sem Park, et al. 2024b. EmbSum: Leveraging the Summarization Capabilities of Large Language Models for Content-Based Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 1010–1015.

Guangping Zhang, Peng Zhang, Jiahao Liu, Zhuoheng Li, Dongsheng Li, Hansu Gu, Tun Lu, and Ning Gu. 2025a. EvalAgent: Towards Evaluating News Recommender Systems with LLM-based Agents. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4086–4095.

Han Zhang, Ziwei Zhu, and James Caverlee. 2023. Evolution of Filter Bubbles and Polarization in News Recommendation. In *European Conference on Information Retrieval*, pages 685–693. Springer.

Lemei Zhang, Peng Liu, and Jon Atle Gulla. 2019. Dynamic Attention-integrated Neural Network for Session-based News Recommendation. *Machine Learning*, 108(10):1851–1875.

Mi Zhang and Neil Hurley. 2008. Avoiding Monotony: Improving the Diversity of Recommendation Lists. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 123–130.

Qi Zhang, Qinglin Jia, Chuyuan Wang, Jingjie Li, Zhaowei Wang, and Xiuqiang He. 2021a. AMM: Attentive Multi-field Matching for News Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1588–1592.

Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021b. UNBERT: User-News Matching BERT for News Recommendation. In *IJCAI*, volume 21, pages 3356–3362.

Shuo Zhang and Peter J Kuhn. 2024. Measuring Bias in Job Recommender Systems: Auditing the Algorithms. Technical report, National Bureau of Economic Research.

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, et al. 2025b. Cold-Start Recommendation towards the Era of Large Language Models (LLMs): A Comprehensive Survey and Roadmap. *arXiv preprint arXiv:2501.01945*.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2021c. Combining Explicit Entity Graph with Implicit Text Information for News Recommendation. In *Companion Proceedings of the Web Conference 2021*, pages 412–416.

Zizhuo Zhang and Bang Wang. 2023. Prompt Learning for News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 227–237.

Wayne Xin Zhao, Yupeng Hou, Xingyu Pan, Chen Yang, Zeyu Zhang, Zihan Lin, Jingsen Zhang, Shuqing Bian, Jiakai Tang, Wenqi Sun, et al. 2022. RecBole 2.0: Towards a More Up-to-Date Recommendation Library. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4722–4726.

Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *proceedings of the 30th acm international conference on information & knowledge management*, pages 4653–4664.

Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C Aggarwal, and Tyler Derr. 2025a. Fairness and Diversity in Recommender Systems: A Survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–28.

Yuyue Zhao, Jin Huang, David Vos, and Maarten de Rijke. 2025b. Revisiting Language Models in Neural News Recommender Systems. In *European Conference on Information Retrieval*, pages 161–176. Springer.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A Deep Reinforcement Learning Framework for News Recommendation. In *Proceedings of the 2018 World Wide Web Conference*, pages 167–176.

Sashuai Zhou, Weinan Gan, Qijiong Liu, Ke Lei, Jieming Zhu, Hai Huang, Yan Xia, Ruiming Tang, Zhenhua Dong, and Zhou Zhao. 2025a. RecBase: Generative Foundation Model Pretraining for Zero-Shot Recommendation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15598–15610.

Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Yuqi Zhou, Sunhao Dai, Liang Pang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2025b. Exploring the Escalation of Source Bias in User, Data, and Recommender System Feedback Loop. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1676–1686.

Zhihui Zhou, Lilin Zhang, and Ning Yang. 2023. Contrastive Collaborative Filtering for Cold-Start Item Recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 928–937.

Feng Zhu, Yan Wang, Chaochao Chen, Jun Zhou, Longfei Li, and Guanfeng Liu. 2021a. Cross-Domain Recommendation: Challenges, Progress, and Prospects. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 4721–4728.

Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. 2019. DAN: Deep Attention Neural Network for News Recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5973–5980.

Yongchun Zhu, Ruobing Xie, Fuzhen Zhuang, Kaikai Ge, Ying Sun, Xu Zhang, Leyu Lin, and Juan Cao. 2021b. Learning to Warm Up Cold Item Embeddings for Cold-start Recommendation with Meta Scaling and Shifting Networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1167–1176.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32.

Itzhak Ziv, Moshe Unger, and Hilah Geva. 2025. The Impact of LLM-Generated Reviews on Recommender Systems: Textual Shifts, Performance Effects, and Strategic Platform Control. *arXiv preprint arXiv:2601.02362*.

Ethan Zuckerman. 2008. The Polyglot Internet.

APPENDIX A

PUBLISHED RESOURCES

Table A.1 provides an overview of the resources (code, data, models) published in the context of this thesis.

Chapter / Section	Resource	Type	Access
§ 3.3.2	NewsRecLib	Code	GitHub:andreeaiana/newsreclib
Chapter 4	Simplifying NNR	Code	GitHub:andreeaiana/simplifying_nnr
Chapter 6	German News	Code	GitHub:andreeaiana/german-news
	US News	Code	GitHub:andreeaiana/us-news
	NeMig	Data	Zenodo: records/7908392
		Code	GitHub:andreeaiana/nemig_nrs
	GeNeG	Data	Zenodo: records/6039372
		Code	GitHub:andreeaiana/geneg
	GeNeG Benchmarking	Code	GitHub:andreeaiana/geneg_benchmarking
Chapter 7	MANNeR	Code	GitHub:andreeaiana/manner
Chapter 8	xMIND	Data	GitHub:andreeaiana/xMIND HuggingFace:aiana94/xMINDsmall HuggingFace:aiana94/xMINDlarge
Chapter 9	PolyNews	Data	HuggingFace:aiana94/polynews
	PolyNewsParallel	Data	HuggingFace:aiana94/polynews-parallel
	NaSE	Code	GitHub:andreeaiana/nase
		Model	HuggingFace:aiana94/NaSE

Table A.1: Overview of all resources published in the context of this thesis.

APPENDIX B

WIKIPEDIA OVER TIME

Year	ENG	ENG %	Top-10	Top-10 %	Total
2025	49.4M	23.5%	124.0M	58.9%	210.6M
2024	47.7M	23.3%	120.1M	59.1%	204.5M
2023	45.8M	23.3%	116.9M	59.6%	196.2M
2022	43.8M	23.3%	112.7M	60.0%	187.8M
2021	41.9M	23.4%	108.6M	60.1%	178.9M
2020	39.7M	23.5%	103.1M	61.0%	168.9M
2019	37.3M	23.9%	96.3M	61.8%	155.9M
2018	35.2M	24.2%	90.1M	62.0%	145.3M
2017	32.9M	24.8%	82.5M	62.3%	132.6M
2016	30.6M	27.0%	68.0M	59.9%	113.4M
2015	28.3M	27.6%	61.4M	59.9%	102.4M
2014	25.5M	28.1%	55.0M	60.7%	90.6M
2013	23.3M	29.2%	49.4M	62.1%	79.7M
2012	21.1M	32.3%	43.1M	66.0%	65.3M
2011	18.7M	34.2%	37.1M	67.9%	54.6M
2010	16.2M	35.6%	31.7M	69.5%	45.5M
2009	13.3M	36.3%	25.9M	70.4%	36.8M
2008	10.5M	36.8%	20.4M	71.6%	28.5M
2007	6.9M	35.6%	14.2M	72.1%	19.7M
2006	3.8M	34.4%	81.9M	73.3%	11.2M
2005	1.5M	32.4%	3.5M	75.7%	4.7M
2004	526K	36.3%	1.4M	80.5%	1.7M
2003	232K	48.4%	436K	90.8%	480K
2002	113K	72.9%	153K	98.9%	155K
2001	20K	89.2%	22K	99.9%	22K

Table B.1: Distribution of number of Wikipedia pages over time. For each year, we show the number of English Wikipedia pages (ENG), the fraction of English Wikipedia pages (ENG %), the number of pages belonging to the ten largest languages (Top-10), the fraction of the pages belonging to the ten largest languages (Top-10 %), and the total number of pages.

APPENDIX C

NEWSRECLIB LOGGING DETAILS

C.1 Configuration Logging

Listing 2 to Listing 6 illustrate an example of how the configuration of each of the pipeline's components is logged when the training process is initiated.

```yaml
csv:
  _target_: lightning.pytorch.loggers.csv_logs.CSVLogger
  save_dir: /pfs/data5/home/ma/ma_ma/ma_aiana/newsreclib/logs/train/runs/nrms_mi
  ↪ ndsmall_pretrainedemb_celoss_bertsent_s42/2023-07-30_19-28-23
  name: csv
  prefix: ''

wandb:
  _target_: lightning.pytorch.loggers.wandb.WandbLogger
  save_dir: /pfs/data5/home/ma/ma_ma/ma_aiana/newsreclib/logs/train/runs/nrms_mi
  ↪ ndsmall_pretrainedemb_celoss_bertsent_s42/2023-07-30_19-28-23
  offline: false
  id: null
  anonymous: null
  project: newsreclib
  log_model: false
  prefix: ''
  group: mind
  tags:
    - nrms
    - mindsmall
    - pretrainedemb
    - celoss
    - bertsent

job_type:
  name: nrms_mindsmall_pretrainedemb_celoss_bertsent_s42
```

Listing 2: Example for logging the configurations of the logger in NewsRecLib.

```yaml
_target_: newsreclib.data.mind_rec_datamodule.MINDRecDataModule
dataset_size: small
dataset_url:
  large:
    train:
    ↪ https://mind201910small.blob.core.windows.net/release/MINDlarge_train.zip
    dev: https://mind201910small.blob.core.windows.net/release/MINDlarge_dev.zip
  small:
    train:
    ↪ https://mind201910small.blob.core.windows.net/release/MINDsmall_train.zip
    dev: https://mind201910small.blob.core.windows.net/release/MINDsmall_dev.zip
pretrained_embeddings_url: https://nlp.stanford.edu/data/glove.840B.300d.zip
data_dir: /pfs/data5/home/ma/ma_ma_aiana/newsreclib/data/
word_embeddings_dirname: glove
word_embeddings_path:
↪ /pfs/data5/home/ma/ma_ma_aiana/newsreclib/data/glove/glove.840B.300d.txt
entity_embeddings_filename: entity_embedding.vec
dataset_attributes:
  - title
  - abstract
  - category
  - subcategory
  - title_entities
  - abstract_entities
  - category_class
  - subcategory_class
  - sentiment_class
  - sentiment_score
id2index_filenames:
  word2index: word2index.tsv
  entity2index: entity2index.tsv
  categ2index: categ2index.tsv
  subcateg2index: subcateg2index.tsv
  sentiment2index: sentiment2index.tsv
  uid2index: uid2index.tsv
use_plm: false
use_pretrained_categ_embeddings: true
word_embed_dim: 300
categ_embed_dim: 300
entity_embed_dim: 100
entity_freq_threshold: 2
entity_conf_threshold: 0.5
sentiment_annotator:
  _target_: newsreclib.data.components.sentiment_annotator.BERTSentimentAnnotator
  model_name: cardiffnlp/twitter-xlm-roberta-base-sentiment
  tokenizer_max_len: 96
valid_time_split: "2019-11-14 00:00:00"
neg_sampling_ratio: 4
max_title_len: 30
max_abstract_len: 50
max_history_len: 50
concatenate_inputs: false
tokenizer_name: null
tokenizer_use_fast: null
tokenizer_max_len: null
batch_size: 64
```

Listing 3: Example for logging the configurations of the data module in NewsRecLib.

```yaml
_target_: newsreclib.models.general_rec.nrms_module.NRMSModule
dual_loss_training: false
dual_loss_coef: null
loss: cross_entropy_loss
late_fusion: false
temperature: None
dataset_attributes:
  - title
  - abstract
  - category
  - subcategory
  - title_entities
  - abstract_entities
  - category_class
  - subcategory_class
  - sentiment_class
  - sentiment_score
attributes2encode:
  - title
use_plm: false
pretrained_embeddings_path: /pfs/data5/home/ma/ma_ma_aiana/newsreclib/data/MINDs
↪ mall_train/transformed_word_embeddings.npy
plm_frozen: null
frozen_layers: null
embed_dim: 300
num_heads: 15
query_dim: 200
dropout_probability: 0.2
outputs:
  train:
    - preds
    - targets
    - cand_news_size
  val:
    - preds
    - targets
    - cand_news_size
  test:
    - preds
    - targets
    - cand_news_size
    - hist_news_size
    - target_categories
    - target_sentiments
    - hist_categories
    - hist_sentiments
num_categ_classes: 18
num_sent_classes: 3
optimizer:
  _target_: torch.optim.Adam
  _partial_: true
  lr: 0.0001
scheduler: null
```

Listing 4: Example for logging the configurations of the recommendation module in NewsRecLib.

```yaml
model_checkpoint:
  _target_: lightning.pytorch.callbacks.ModelCheckpoint
  dirpath: /pfs/data5/home/ma/ma_ma/ma_aiana/newsreclib/logs/train/runs/nrms_min
  ↪ dsmall_pretrainedemb_celoss_bertsent_s42/2023-07-30_19-13-45/checkpoints
  filename: epoch{epoch:03d}
  monitor: val/loss
  verbose: false
  save_last: true
  save_top_k: 1
  mode: min
  auto_insert_metric_name: false
  save_weights_only: false
  every_n_train_steps: null
  train_time_interval: null
  every_n_epochs: null
  save_on_train_epoch_end: null

early_stopping:
  _target_: lightning.pytorch.callbacks.EarlyStopping
  monitor: val/loss
  min_delta: 0.0
  patience: 5
  verbose: false
  mode: max
  strict: true
  check_finite: true
  stopping_threshold: null
  divergence_threshold: null
  check_on_train_epoch_end: null

model_summary:
  _target_: lightning.pytorch.callbacks.RichModelSummary
  max_depth: -1

rich_progress_bar:
  _target_: lightning.pytorch.callbacks.RichProgressBar
```

Listing 5: Example for logging the configurations of the callbacks in NewsRecLib.

```yaml
_target_: lightning.pytorch.trainer.Trainer
default_root_dir: /pfs/data5/home/ma/ma_ma/ma_aiana/newsreclib/logs/train/runs/n
↪ rms_mindsmall_pretrainedemb_celoss_bertsent_s42/2023-07-30_19-28-23
min_epochs: 1
max_epochs: 10
accelerator: gpu
devices: 1
precision: 16
check_val_every_n_epoch: 1
deterministic: false
profiler: simple
```

Listing 6: Example for logging the configurations of the trainer in NewsRecLib.

C.2 Model Metadata Logging

Fig. C.1 shows an example of logging relevant metadata information regarding a model's size and number of parameters.

	Name	Type	Params
0	train_loss	MeanMetric	0
1	val_loss	MeanMetric	0
2	test_loss	MeanMetric	0
3	val_loss_best	MinMetric	0
4	criterion	CrossEntropyLoss	0
5	news_encoder	NewsEncoder	16.9 M
6	news_encoder.text_encoders	ModuleDict	16.9 M
7	news_encoder.text_encoders.title	MHSAAddAtt	16.9 M
8	news_encoder.text_encoders.title.embedding_layer	Embedding	16.5 M
9	news_encoder.text_encoders.title.multihead_attention	MultiheadAttention	361 K
10	news_encoder.text_encoders.title.multihead_attention.out_proj	NonDynamicallyQuantizableLinear	90.3 K
11	news_encoder.text_encoders.title.additive_attention	AdditiveAttention	60.4 K
12	news_encoder.text_encoders.title.additive_attention.linear	Linear	60.2 K
13	news_encoder.text_encoders.title.dropout	Dropout	0
14	user_encoder	UserEncoder	421 K
15	user_encoder.multihead_attention	MultiheadAttention	361 K
16	user_encoder.multihead_attention.out_proj	NonDynamicallyQuantizableLinear	90.3 K
17	user_encoder.additive_attention	AdditiveAttention	60.4 K
18	user_encoder.additive_attention.linear	Linear	60.2 K
19	click_predictor	DotProduct	0
20	train_rec_metrics	MetricCollection	0
21	train_rec_metrics.auc	BinaryAUROC	0
22	train_rec_metrics.mrr	RetrievalMRR	0
23	train_rec_metrics.ndcg@10	RetrievalNormalizedDCG	0
24	train_rec_metrics.ndcg@5	RetrievalNormalizedDCG	0
25	val_rec_metrics	MetricCollection	0
26	val_rec_metrics.auc	BinaryAUROC	0
27	val_rec_metrics.mrr	RetrievalMRR	0
28	val_rec_metrics.ndcg@10	RetrievalNormalizedDCG	0
29	val_rec_metrics.ndcg@5	RetrievalNormalizedDCG	0
30	test_rec_metrics	MetricCollection	0
31	test_rec_metrics.auc	BinaryAUROC	0
32	test_rec_metrics.mrr	RetrievalMRR	0
33	test_rec_metrics.ndcg@10	RetrievalNormalizedDCG	0
34	test_rec_metrics.ndcg@5	RetrievalNormalizedDCG	0
35	test_categ_div_metrics	MetricCollection	0
36	test_categ_div_metrics.categ_div@10	Diversity	0
37	test_categ_div_metrics.categ_div@5	Diversity	0
38	test_sent_div_metrics	MetricCollection	0
39	test_sent_div_metrics.sent_div@10	Diversity	0
40	test_sent_div_metrics.sent_div@5	Diversity	0
41	test_categ_pers_metrics	MetricCollection	0
42	test_categ_pers_metrics.categ_pers@10	Personalization	0
43	test_categ_pers_metrics.categ_pers@5	Personalization	0
44	test_sent_pers_metrics	MetricCollection	0
45	test_sent_pers_metrics.sent_pers@10	Personalization	0
46	test_sent_pers_metrics.sent_pers@5	Personalization	0

Trainable params: 17.4 M
Non-trainable params: 0
Total params: 17.4 M
Total estimated model params size (MB): 69

Figure C.1: Example for logging the model size, number of trainable and non-trainable model parameters in NewsRecLib.

APPENDIX D

ADDITIONAL RESULTS FOR CHAPTER 4

Fig. D.1 shows the number of trainable parameters in original EF configurations, on MINDlarge.

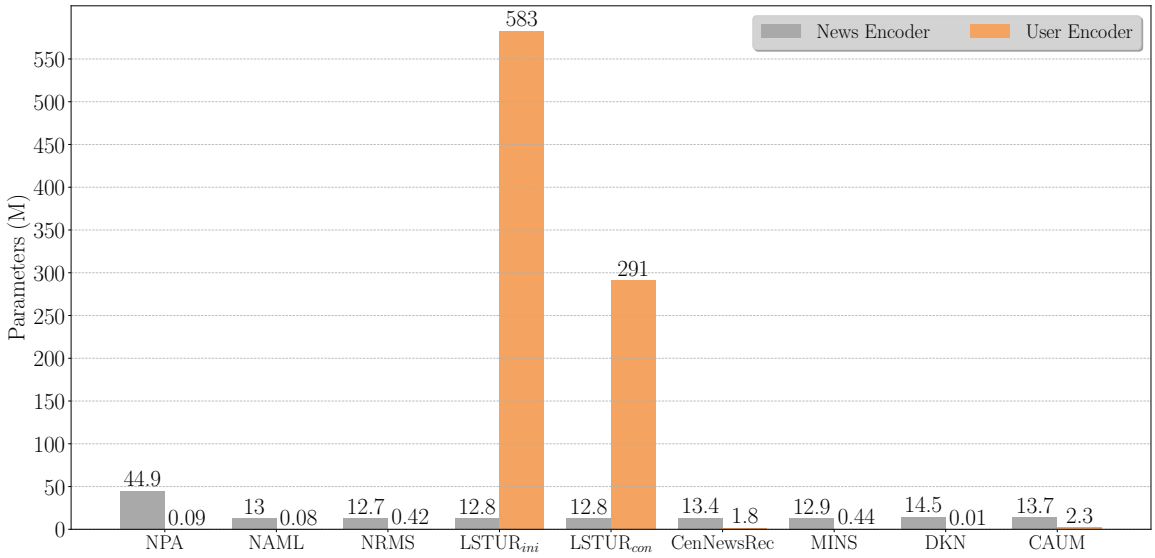


Figure D.1: Number of model parameters (in millions) on MINDlarge.

APPENDIX E

EXPERIMENTAL DETAILS FOR CHAPTER 7

E.1 Reproducibility

E.1.1 Model Parameters

		MINDlarge		Adressa	
Model	Non-trainable	Trainable	Total	Trainable	Total
NRMS-PLM	56.7	73	129	126	182
MINER	56.7	68.2	124	121	178
NAML-PLM	56.7	70.8	127	124	180
LSTUR-PLM	56.7	633	690	200	257
MINS-PLM	56.7	73.3	130	126	183
$\text{CAUM}_{\text{no entities}}$-PLM	56.7	73.2	129	126	183
CAUM-PLM	56.7	74.9	131	–	–
TANR-PLM	56.7	70.6	127	123	180
SentiRec-PLM	56.7	73	129	126	182
SentiDebias-PLM	56.7	73.3	130	126	183
MANNeR ($\text{CR-Module}_{\text{title}}$ / $\text{A-Module}_{\text{title}}$) – monolingual	56.7	67.9	124	121	177
MANNeR (CR-Module/ A-Module) – monolingual	56.7	70.3	126	–	–
MANNeR ($\text{CR-Module}_{\text{title}}$ / $\text{A-Module}_{\text{title}}$) – multilingual	0	134	134	134	134

Table E.1: Number of model parameters (in millions). $\text{CR-Module}_{\text{title}}$ / $\text{A-Module}_{\text{title}}$ denote the MANNeR modules trained with only the news title as input to the NE.

E.1.2 Hyperparameters and Implementation

Hyperparameter Optimization. We use $\text{RoBERTa}_{\text{Base}}$ (Liu et al., 2019b) and $\text{NB-BERT}_{\text{Base}}$ (Kummervold et al., 2021; Nielsen, 2023) as the backbone PLMs of all models, in experiments on MIND and Adressa, respectively. In both cases, we fine-tune only the PLM's last four layers.[1] In the cross-lingual transfer experiments from § 7.5.4, we fine-tune all of the 6 layers of DistilBERT. We use 100-dimensional TransE embeddings (Bordes et al., 2013) pretrained on Wikidata as input to the entity encoder in the NE of the knowledge-aware NNRs. We perform hyperparameter tuning on the main hyperparameters ofMANNeR and the baselines using grid search. Table E.2 lists the search spaces for all the optimized hyperparameters, as well as the best values. We repeat each experiment five times with the seeds ($\{42, 43, 44, 45, 46\}$) set with PyTorch Lightning's `seed_everything`.

[1] In preliminary results, we did not see significant differences between full fine-tuning of all layers and fine-tuning only the last four layers. In the interest of computational efficiency, we thus froze the first eight layers of the Transformer.

	lr	$\text{num}_{\text{heads}}$	$\text{query}_{\text{dim}}$	UE	K	score agg	λ	μ	α	$\tau_{CR-Module}$	$\tau_{A-Module}$
Search Space	$[1e^{-4}, 1e^{-6}]$	{8, 12, 16, 24, 32}	[50, 200]	{ini, con}	{8, 16, 32, 48}	{mean, max, weighted}	[0.1, 0.3]	[5, 15]	[0.05, 0.2]	[0.1, 0.5]	[0.1, 0.9]
Step	$1e^{-1}$	–	50	–	–	–	0.05	5	0.05	0.02	0.05
NRMS-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 8	150 / 200	–	–	–	–	–	–	–	–
MINER	$1e^{-5}$ / $1e^{-6}$	–	–	–	32 / 48	mean / mean	–	–	–	–	–
NAML-PLM	$1e^{-5}$ / $1e^{-6}$	16 / 8	200 / 200	–	–	–	–	–	–	–	–
LSTUR-PLM	$1e^{-5}$ / $1e^{-6}$	24 / 8	150 / 50	ini / ini	–	–	–	–	–	–	–
MINS-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 12	100 / 200	–	–	–	–	–	–	–	–
CAUM-PLM	$1e^{-5}$ / $1e^{-6}$	16 / 16	50 / 150	–	–	–	–	–	–	–	–
TANR-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 8	150 / 50	–	–	–	0.3 / 0.15	–	–	–	–
SentiRec-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 8	200 / 200	–	–	–	–	5 / 5	–	–	–
SentiDebias-PLM	$1e^{-5}$ / $1e^{-6}$	8 / 12	100 / 150	–	–	–	–	–	0.15 / 0.15	–	–
MANNeR	$1e^{-5}$ / $1e^{-6}$	–	200 / 200	–	–	–	–	–	–	0.36 / 0.14	0.9 / 0.9

Table E.2: Search spaces used for hyperparameter optimization and best values found for all models. We report the optimal values in the format $value_{MINDlarge}$ / $value_{Adressa}$. We use the following abbreviations: lr = learning rate, $\text{num}_{\text{heads}}$ = number of attention heads, $\text{query}_{\text{dim}}$ = dimensionality of the query vector in additive attention, UE agg = aggregation method used to combine the long-term and the short-term user representations into a final user embedding in LSTUR (An et al., 2019), K = number of context codes in MINER (Li et al., 2022), score agg = aggregation function for the final user click score calculation in MINER (Li et al., 2022), λ = weight of the topic classification task in TANR (Wu et al., 2019c), μ = weight of the sentiment diversity regularization loss in SentiRec (Wu et al., 2020c), α = adversarial loss coefficient in SentiDebias (Wu et al., 2022h), τ = temperature parameter in SCL in MANNeR, *ini* = initialize, *con* = concatenate, *categ* = category.

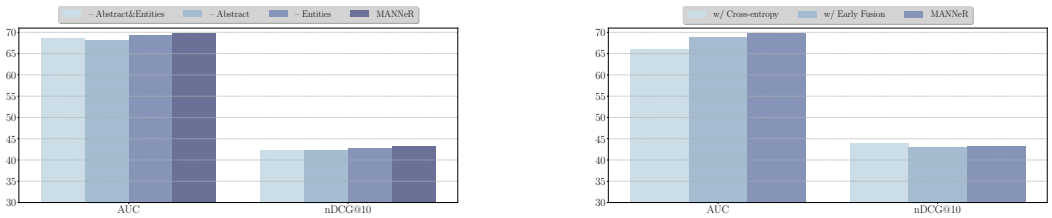
Code. We train MANNeR, as well as all the baselines, using the implementations provided in the NewsRecLib library (see § 3.3.2).

Infrastructure and Compute. We conduct all experiments on a cluster with virtual machines. We train MANNeR on both datasets, as well as the baselines on MIND, on a single NVIDIA A100 40GB GPU. We train the baselines on Adressa on a single NVIDIA Tesla V100 32GB GPU.

E.2 Additional Results

E.2.1 Content Personalization

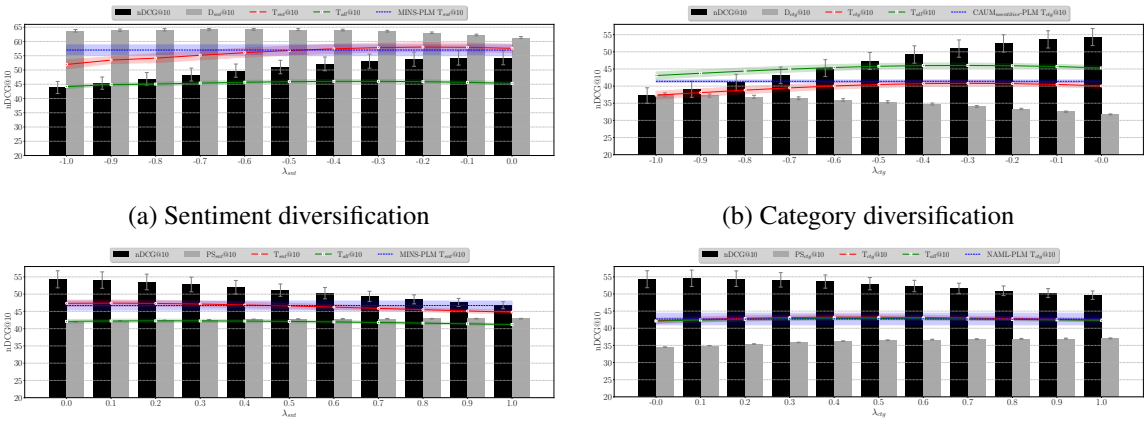
Fig. E.1a shows MANNeR's performance on MIND for different inputs to the NE. Even the CR-Module exposed to titles only (i.e., no abstract or entity information) outperforms all of the baselines on content recommendation. Fig. E.1b illustrates MANNeR's performance for alternative architecture designs and training objectives (cf. § 7.5.1).[2]



(a) Input features for the news encoder (NE)

(b) CR-Module design/training alternatives

Figure E.1: Effect of different (a) NE inputs and (b) model design/training choices on MANNeR's content-based personalization performance.



(a) Sentiment diversification

(b) Category diversification

(c) Sentiment personalization

(d) Category personalization

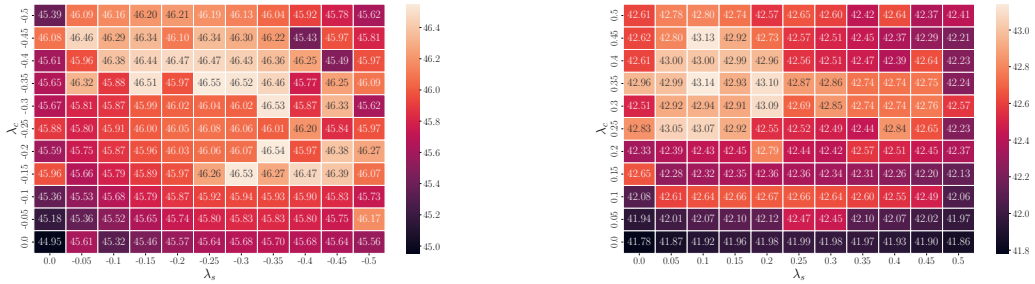
Figure E.2: Results of single-aspect customization for MANNeR and the best baseline on Adressa.

E.2.2 Single-Aspect Customization

Fig. E.2 explores the trade-off between content and aspect diversification, and respectively, personalization tasks for different values of λ_{ctg} and λ_{snt} on the Adressa dataset.

E.2.3 Multi-Aspect Customization

Fig. E.3 explores the trade-off between content personalization and multi-aspect diversification on Adressa.



(a) Multi-aspect diversification on Adressa

(b) Multi-aspect personalization on Adressa

Figure E.3: Results of multi-aspect customization for MANNeR on Adressa.

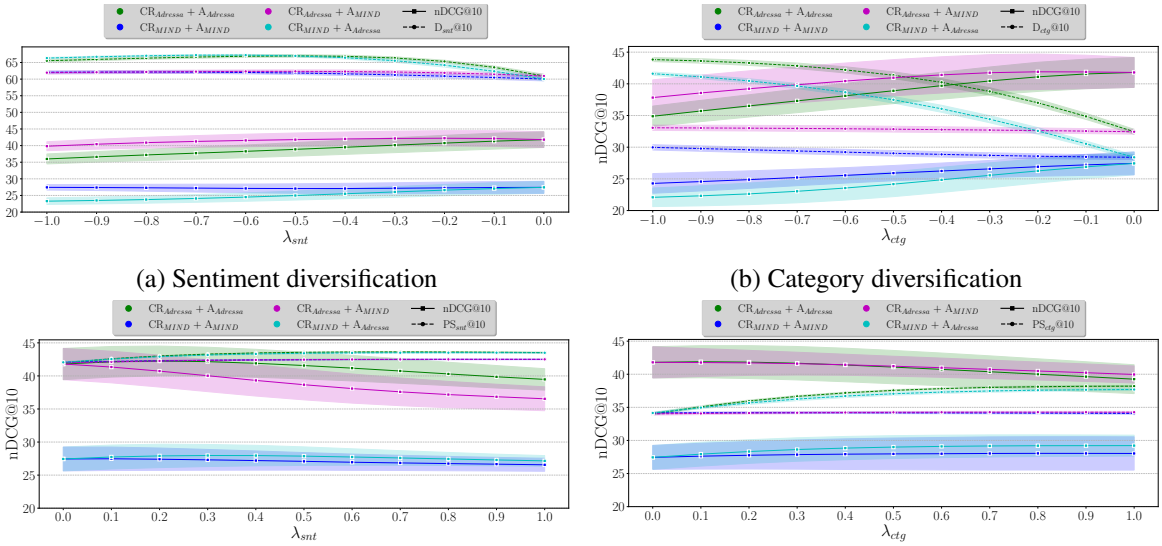
E.2.4 Cross-Lingual Transfer

Fig. E.4 summarizes the XLT results for single-aspect diversification and personalization, respectively, on the target-language dataset Adressa.

E.2.5 Time Complexity Analysis

Table E.3 shows the average inference time for the entire MIND (365,201 impressions), and respectively, Adressa (146,284 impressions) test sets. Note that runtimes heavily depend on the computing infrastructure used, as well as on the parallel usage of the infrastructure for other tasks, as experiments

[2]For brevity, we report results on MIND; findings on Adressa exhibit identical trends.



(a) Sentiment diversification

(b) Category diversification

(c) Sentiment personalization

(d) Category personalization

Figure E.4: XLT in single-aspect diversification and personalization, with modules trained on different (combinations of) source-language datasets and evaluated on the target-language dataset Adressa. The line style indicates the metric, the color the source-language datasets used in training.

are conducted on a HPC cluster. We highlight that MANNeR achieves a much lower inference time than the other NNRs.

Model	**MINDlarge**	**Adressa**
NRMS-PLM	$17.53_{\pm 0.48}$	$7.13_{\pm 0.27}$
MINER	$16.03_{\pm 1.66}$	$9.96_{\pm 0.73}$
NAML-PLM	$33.99_{\pm 0.51}$	$7.09_{\pm 0.14}$
MINS-PLM	$27.50_{\pm 10.87}$	$7.81_{\pm 0.21}$
$\text{CAUM}_{\text{no entities}}$-PLM	$22.67_{\pm 2.46}$	$8.12_{\pm 0.13}$
CAUM-PLM	$25.22_{\pm 0.45}$	–
TANR-PLM	$17.02_{\pm 1.07}$	$6.98_{\pm 0.08}$
SentiRec-PLM	$17.93_{\pm 0.34}$	$7.02_{\pm 0.08}$
SentiDebias-PLM	$21.01_{\pm 3.03}$	$13.28_{\pm 0.83}$
MANNeR (CR-Module)	$1.34_{\pm 0.03}$	$2.09_{\pm 0.06}$
MANNeR (CR-Module + *ctg* A-Module)	$1.68_{\pm 0.08}$	$2.78_{\pm 0.10}$
MANNeR (CR-Module + *snt* A-Module)	$1.65_{\pm 0.01}$	$2.73_{\pm 0.06}$
MANNeR (CR-Module + 2 A-Module)	$2.13_{\pm 0.05}$	$3.17_{\pm 0.01}$

Table E.3: Inference time (in thousands of seconds) for the different NNRs on the test portions of the MINDlarge and Adressa datasets, respectively.

APPENDIX F

EXPERIMENTAL DETAILS FOR CHAPTER 8

F.1 Language Model Support

Table F.1 indicates whether the languages from xMIND are included in the pretraining corpora of the mPLMs used in our experiments (cf. Section 8.4).

Language Code	SWH	SOM	CMN	JPN	TUR	TAM	VIE	THA	RON	FIN	KAT	HAT	IND	GRN
XLM-RoBERTa (Conneau et al., 2020)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
LaBSE (Feng et al., 2022)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
umT5 (Chung et al., 2023)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

Table F.1: Support of the 14 languages of xMIND by different mPLMs.

F.2 Hyperparameter Optimization

	$\texttt{num}_{\texttt{heads}}$	$\texttt{query}_{\texttt{dim}}$	K	score agg	τ
Search Space	{8, 12, 16, 24, 32}	[50, 200]	{8, 16, 32, 48}	{mean, max, weighted}	[0.1, 0.5]
Step	–	–	50	–	0.02
CAUM	8 / 8 / 8	200 / 200 / 200	–	–	–
LSTUR	16 / 8 / 8	150 / 50 / 200	–	–	–
MANNeR	–	–	–	–	0.38 / 0.38 / 0.38
MINER	–	–	8 / 32 / 48	mean / mean / weighted	–
MINS	32 / 16 / 32	50 / 100 / 200	–	–	–
NAML	24 / 24 / 8	200 / 50 / 200	–	–	–
TANR	24 / 12 / 16	150 / 200 / 100	–	–	–

Table F.2: Search spaces used for hyperparameter optimization and best values found for all models. We report the optimal values depending on the underlying language model, in the format $value_{XLM\text{-}RoBERTa}$ / $value_{LaBSE}$ / $value_{umT5}$. We use the following abbreviations: $\texttt{num}_{\texttt{heads}}$ = number of attention heads, $\texttt{query}_{\texttt{dim}}$ = dimensionality of the query vector in additive attention, K = number of context codes in MINER (Li et al., 2022), score agg = aggregation function for the final user click score calculation in MINER (Li et al., 2022), τ = temperature parameter in supervised contrastive loss in MANNeR (see Chapter 7).

We perform hyperparameter optimization for the most important hyperparameters of each NNR using grid search and the English news as training and validation sets (i.e., hyperparameter tuning on

the MIND dataset). Concretely, we search for the optimal learning rate in the range $[1e^{-3}, 1e^{-4}, 1e^{-5}]$ for all models, and find $1e^{-5}$ to be the most suitable value for all NNRs. Table F.2 lists the search spaces for the other tuned hyperparameters, as well as the best values for each model. We set all remaining model-specific hyperparameters to the optimal values reported in the respective papers. We repeat each experiment three times with different random seeds (i.e., {42, 43, 44}) set with PyTorch Lightning's `set_everything`.

APPENDIX G

EXPERIMENTAL DETAILS FOR CHAPTER 9

G.1 PolyNews Corpora

G.1.1 Language Characteristics

Table G.1 lists the languages included in the PolyNews dataset, summarizing the following information, according to the #BenderRule:

- **Code:** We denote each language using a BCP 47 tag sequence that combines a base subtag indicating the three-letter ISO 693-3 code with the ISO 15924 script subtag. We use the script code to differentiate between languages collected in multiple scripts.
- **Language:** In case of multiple denominations, we use the language name from Ethnologue (Lewis et al., 2009), cross-referenced against other linguistic resources, namely Glottolog (Hammarström et al., 2021) and World Atlas of Structures (Dryer and Haspelmath, 2013).
- **Script:** We provide the English name of the script.
- **Macro-area:** We indicate the macro-area as per WALS (Dryer and Haspelmath, 2013).
- **Family and Subgrouping:** We specify the language family and subgrouping from WALS (Dryer and Haspelmath, 2013) and Glottolog (Hammarström et al., 2021).
- **Total Speakers:** We report the total number of speakers of the language, considering both L1-level (first-language) and L2-level (second-language) speakers, according to Ethnologue.[1]
- **LaBSE support:** We indicate whether the language was included in the pretraining corpora of LaBSE (Feng et al., 2022).
- **Average byte length:** We report the average number of bytes per text for each language.
- **Average character length:** We report the average number of characters per text for each language.

[1]We use the statistics available in May 2024 at https://www.ethnologue.com/.

Code	Language	Script	Macro-area	Family	Subgrouping	# Speakers (M)	LaBSE	Avg. # bytes	Avg. # chars.
amh_Ethi	**Amharic**	Ethiopic	Africa	Afro-Asiatic	Semitic	59.69	✓	190.42	80.10
arb_Arab	**Modern Standard Arabic**	Arabic	Eurasia	Afro-Asiatic	Semitic	332.46	✓	225.86	130.60
ayr_Latn	**Central Aymara**	Latin	South America	Aymaran	Central Southern Aymara	1.46	✗	122.33	118.48
bam_Latn	**Bambara**	Latin	Africa	Mande	Western Mande	14.19	✗	135.94	122.92
bbj_Latn	**Ghomálá'**	Latin	Africa	Niger-Congo	Atlantic-Congo	0.35	✗	86.77	64.87
ben_Beng	**Bengali**	Bengali	Eurasia	Indo-European	Indo-Aryan	278.18	✓	300.72	128.01
bos_Latn	**Bosnian**	Latin	Eurasia	Indo-European	Balto-Slavic	2.68	✓	52.08	50.76
bul_Cyrl	**Bulgarian**	Cyrillic	Eurasia	Indo-European	Balto-Slavic	7.88	✓	256.64	148.65
cat_Latn	**Catalan**	Latin	Eurasia	Indo-European	Italic	9.29	✓	132.07	129.25
ces_Latn	**Czech**	Latin	Eurasia	Indo-European	Balto-Slavic	12.32	✓	127.28	114.53
ckb_Arab	**Central Kurdish**	Arabic	Eurasia	Indo-European	Indo-Iranian	5.35	✓	350.93	185.93
dan_Latn	**Danish**	Latin	Eurasia	Indo-European	Germanic	5.81	✓	133.51	130.71
deu_Latn	**German**	Latin	Eurasia	Indo-European	Germanic	133.91	✓	136.53	134.47
ell_Grek	**Greek**	Greek	Eurasia	Indo-European	Greek	13.23	✓	273.36	159.73
eng_Latn	**English**	Latin	Eurasia	Indo-European	Germanic	1,515.23	✓	123.89	123.28
est_Latn	**Estonian**	Latin	Eurasia	Uralic	Finnic	1.09 (L1 only)	✓	116.43	113.24
ewe_Latn	**Éwé**	Latin	Africa	Niger-Congo	Atlantic-Congo	5.53	✗	78.44	72.39
fil_Latn	**Filipino**	Latin	Papunesia	Austronesian	Malayo-Polynesian	–	✗	135.54	135.24
fin_Latn	**Finnish**	Latin	Eurasia	Uralic	Finnic	5.61	✓	117.52	112.76
fon_Latn	**Fon**	Latin	Africa	Niger-Congo	Atlantic-Congo	2.28	✗	179.91	136.43
fra_Latn	**French**	Latin	Eurasia	Indo-European	Italic	311.58	✓	143.82	139.25
guj_Gujr	**Gujarati**	Gujarati	Eurasia	Indo-European	Indo-Iranian	62.67	✓	301.16	118.81
guw_Latn	**Gun**	Latin	Africa	Niger-Congo	Atlantic-Congo	1.54	✗	93.61	76.95
hau_Latn	**Hausa**	Latin	Africa	Afro-Asiatic	Chadic	88.24	✓	138.36	137.40
heb_Hebr	**Hebrew**	Hebrew	Eurasia	Afro-Asiatic	Semitic	9.36	✓	132.72	77.33
hin_Deva	**Hindi**	Devanagari	Eurasia	Indo-European	Indo-Iranian	608.85	✓	321.32	129.44
hun_Latn	**Hungarian**	Latin	Eurasia	Uralic	–	12.43	✓	141.26	128.87
ibo_Latn	**Igbo**	Latin	Eurasia	Niger-Congo	Attlantic-Congo	30.91	✓	123.64	104.73
ind_Latn	**Indonesian**	Latin	Papunesia	Austronesian	Malayo-Polynesian	199.00	✓	131.97	131.56
ita_Latn	**Italian**	Latin	Eurasia	Indo-European	Italic	66.79	✓	138.79	137.64
jpn_Jpan	**Japanese**	Japanese	Eurasia	Japonic	Japanesic	123.47	✓	160.54	64.63
kaz_Cyrl	**Kazakh**	Cyrillic	Eurasia	Altaic	Turkic	16.58	✓	279.77	155.47
khm_Khmr	**Khmer**	Khmer	Eurasia	Austro-Asiatic	Khmeric	17.60	✓	368.65	139.12
kor_Hang	**Korean**	Hangul	Eurasia	Koreanic	Korean	81.13	✓	156.69	72.81
lav_Latn	**Latvian**	Latin	Eurasia	Indo-European	Baltic	1.76	✓	135.63	124.39
lin_Latn	**Lingala**	Latin	Africa	Niger-Congo	Atlantic-Congo	40.54	✗	66.14	65.62
lit_Latn	**Lithuanian**	Latin	Eurasia	Indo-European	Balto-Slavic	2.78	✓	135.97	126.83
lug_Latn	**Ganda**	Latin	Africa	Niger-Congo	Atlantic-Congo	11.14	✗	111.21	110.90
luo_Latn	**Luo**	Latin	Africa	Nilotic	Western Nilotic	5.26	✗	158.18	157.76
mkd_Cyrl	**Macedonian**	Cyrillic	Eurasia	Indo-European	Balto-Slavic	1.65	✓	261.49	150.68
mos_Latn	**Mossi**	Latin	Africa	Niger-Congo	Atlantic-Congo	11.87	✗	140.87	127.62
mya_Mymr	**Burmese**	Myanmar	Eurasia	Sino-Tibetan	Tibeto-Burman	43.17	✓	383.44	148.70
nld_Latn	**Dutch**	Latin	Eurasia	Indo-European	Germanic	25.31	✓	124.46	123.98
nor_Latn	**Norwegian**	Latin	Eurasia	Indo-European	Germanic	5.42	✓	44.14	43.20
npi_Deva	**Nepali**	Devanagari	Eurasia	Indo-European	Indo-Iranian	31.98	✓	441.96	178.46
orm_Latn	**Oromo**	Latin	Africa	Afro-Asiatic	Cushitic	45.53 (L1 only)	✗	79.86	79.57
ory_Orya	**Odia**	Oriya	Eurasia	Indo-European	Indo-Iranian	39.76	✓	268.32	107.68
pan_Guru	**Eastern Panjabi**	Gurmukhi	Eurasia	Indo-European	Indo-Iranian	31.17	✓	685.89	275.91
pcm_Latn	**Nigerian Pidgin**	Latin	Africa	Creole	–	120.65	✗	118.72	118.69
pes_Arab	**Western Persian**	Arabic	Eurasia	Indo-European	Indo-Iranian	78.83	✓	218.93	128.05
plt_Latn	**Malagasy**	Latin	Africa	Austronesia	Malayo-Polinesian	7.55	✓	152.57	151.77
pol_Latn	**Polish**	Latin	Eurasia	Indo-European	Balto-Slavic	41.80	✓	120.85	114.56
por_Latn	**Portuguese**	Latin	Eurasia	Indo-European	Italic	263.84	✓	124.35	120.73
ron_Latn	**Romanian**	Latin	Eurasia	Indo-European	Italic	25.23	✓	129.89	123.63
run_Latn	**Rundi**	Latin	Africa	Niger-Congo	Atlantic-Congo	13.19	✗	75.73	75.20
rus_Cyrl	**Russian**	Cyrillic	Eurasia	Indo-European	Balto-Slavic	255.39	✓	151.75	88.92
sna_Latn	**Shona**	Latin	Africa	Niger-Congo	Atlantic-Congo	10.88	✓	67.94	67.77
som_Latn	**Somali**	Latin	Africa	Afro-Asiatic	Cushitic	23.88	✓	73.09	72.98
spa_Latn	**Spanish**	Latin	Eurasia	Indo-European	Italic	559.52	✓	136.71	134.00
sqi_Latn	**Albanian**	Latin	Eurasia	Indo-European	Albanian	6.39	✓	128.83	119.94
srp_Cyrl	**Serbian**	Cyrillic	Eurasia	Indo-European	Balto-Slavic	10.12	✓	88.89	51.97
srp_Latn	**Serbian**	Latin	Eurasia	Indo-European	Balto-Slavic	10.12	✓	78.22	76.17
swe_Latn	**Swedish**	Latin	Eurasia	Indo-European	Germanic	13.25	✓	112.77	108.45
swh_Latn	**Swahili**	Latin	Africa	Niger-Congo	Atlantic-Congo	86.52	✓	115.32	115.00
tam_Taml	**Tamil**	Tamil	Eurasia	Dravidian	Southern Dravidian	86.72	✓	170.92	64.02
tet_Latn	**Tetun**	Latin	Africa	Austronesian	Malayo-Polynesian	0.49	✗	130.19	128.88
tha_Thai	**Thai**	Thai	Eurasia	Kra-Dai	Kam-Tai	61.10	✓	144.90	56.57
tir_Ethi	**Tigrinya**	Ethiopic	Africa	Afro-Asiatic	Semitic	9.87	✗	116.47	47.10
tsn_Latn	**Tswana**	Latin	Africa	Niger-Congo	Atlantic-Congo	13.75	✗	142.02	141.69
tur_Latn	**Turkish**	Latin	Eurasia	Altaic	Turkic	90.09	✓	122.87	112.45
twi_Latn	**Twi**	Latin	Africa	Niger-Congo	Atlantic-Congo	0.024	✗	119.62	109.66
ukr_Cyrl	**Ukrainian**	Cyrillic	Eurasia	Indo-European	Balto-Slavic	38.92	✓	107.62	60.99
urd_Arab	**Urdu**	Arabic	Eurasia	Indo-European	Indo-Iranian	169.70	✓	193.15	115.04
wol_Latn	**Wolof**	Latin	Africa	Niger-Congo	Atlantic-Congo	22.65	✓	128.05	122.11
xho_Latn	**Xhosa**	Latin	Africa	Niger-Congo	Atlantic-Congo	19.22	✓	47.79	47.37
yor_Latn	**Yorùbá**	Latin	Africa	Niger-Congo	Atlantic-Congo	47.19	✓	174.46	124.68
zho_Hans	**Chinese**	Han (Simplified)	Eurasia	Sino-Tibetan	Sinitic	1,352.67 (L1 only)	✓	78.06	36.56
zho_Hant	**Chinese**	Han (Traditional)	Eurasia	Sino-Tibetan	Sinitic	1,352.67 (L1 only)	✗	76.86	36.64
zul_Latn	**Zulu**	Latin	Africa	Niger-Congo	Atlantic-Congo	27.80	✓	165.13	164.64

Table G.1: The 77 languages of PolyNews. We display the language *Code* (ISO 693-3), language name, *Script*, *Macro-area*, *Family* and *Genus*, and report the total number of L1-level and L2-level speakers of the language according to Ethnologue (Lewis et al., 2009). The eighth column indicates whether the language was included in the pretraining corpora of LaBSE (Feng et al., 2022). The last two columns specify the average byte and character lengths of the texts for each language. The languages highlighted in gray were not included in the adaptive pretraining of NaSE.

G.1.2 Parallel Corpora Statistics

Fig. G.1 illustrates the number of texts for each language pair available in the PolyNewsParallel dataset. Note that PolyNewsParallel consists of only 64 out of the 77 languages listed in Table G.1.

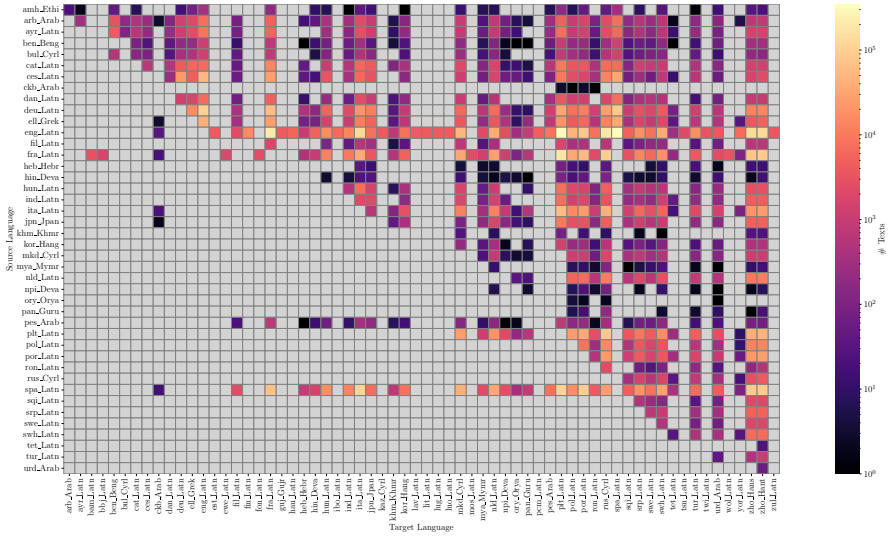


Figure G.1: Distribution of texts across the 833 language pairs in PolyNewsParallel. The gray cells indicate that no texts exist for the given language pair.

G.2 Reproducibility

Hyperparameter Optimization. We perform hyperparameter optimization on the main hyperparameters of all NNRs using grid search. More specifically, we search for the optimal learning rate in $\{1e^{-3}, 1^{e-4}, 1^{e-5}\}$, finding $1e^{-5}$ to be the most suitable value for all NNRs. We optimize the number of heads in the multi-head self-attention networks of NAML, MINS, CAUM in $[8, 12, 16, 24, 32]$, and the query vector dimensionality in the additive attention network of NAML and MINS in $[50, 200]$, with a step of 50. We use the following best performing settings: 32 attention heads for NAML and MINS and 8 for CAUM when coupled with XLM-RoBERTa$_{\text{Large}}$, and 24, 16, and 8 attention heads for NAML, MINS, and CAUM, respectively, when using LaBSE or NaSE as the NE's backbone. Moreover, we use a query vector of dimension 200 for NAML and 50 for MINS in combination with XLM-RoBERTa$_{\text{Large}}$, whereas for NEs based on LaBSE or NaSE we use a dimensionality of 50 for NAML and 100 for MINS. We find the optimal temperature of 0.38 for the supervised contrastive loss in MANNeR and LFRec-SCL by sweeping the interval $[0.1, 0.5]$, with a step of 0.02. We set all the remaining model-specific hyperparameters to the optimal values reported in the respective papers. We repeat each experiment three times with the seed ($\{42, 43, 44\}$) set with PyTorch Lightning's `seed_everything`.

Code. We train NaSE[2] using Sentence Transformers (Reimers and Gurevych, 2019) and PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019).[3] We conduct all news recommendation experiments with the NewsRecLib library (see § 3.3.2).

Infrastructure and Compute. We train NaSE on a cluster with virtual machines, on single NVIDIA A100 40GB GPUs. We conduct the news recommendation experiments on the same cluster, on single NVIDIA A40 48GB GPUs.

[2] https://github.com/andreeaiana/nase
[3] https://lightning.ai/docs/pytorch/stable/

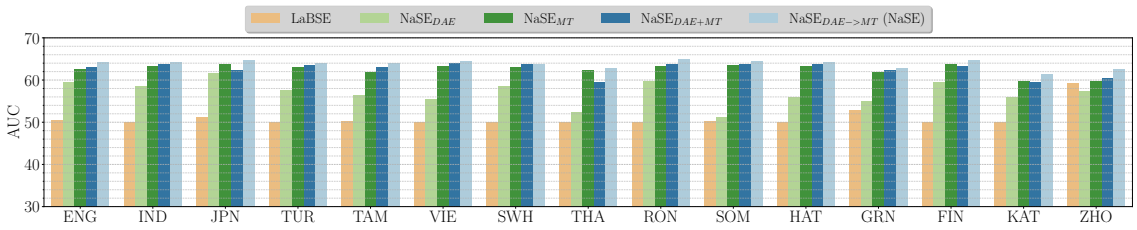


Figure G.2: Recommendation performance (AUC) of LFRec with frozen NaSE embeddings for different training strategies of NaSE versus LFRec with frozen LaBSE embeddings, across English and the 14 languages in xMIND.

G.3 Additional Results

Fig. G.2 illustrates the performance in terms of AUC of LFRec with frozen NaSE and LaBSE embeddings on the MIND and xMIND datasets.