MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation

Andreea Iana
University of Mannheim
Mannheim, Germany
andreea.iana@uni-mannheim.de

Goran Glavaš
University of Würzburg
Würzburg, Germany
goran.glavas@uni-wuerzburg.de

Heiko Paulheim
University of Mannheim
Mannheim, Germany
heiko.paulheim@uni-mannheim.de

ABSTRACT

Digital news platforms use news recommenders as the main instrument to cater to the individual information needs of readers. Despite an increasingly language-diverse online community, in which many Internet users consume news in multiple languages, the majority of news recommendation focuses on major, resource-rich languages. Moreover, nearly all news recommendation efforts assume *monolingual* news consumption, whereas more and more users tend to consume information in at least two languages. Accordingly, the existing body of work on news recommendation suffers from a lack of publicly available multilingual benchmarks that would catalyze development of news recommenders effective in multilingual settings and for low-resource languages. Aiming to fill this gap, we introduce xMIND, an *open, multilingual* news recommendation dataset derived from the English MIND dataset using machine translation, covering a set of 14 linguistically and geographically diverse languages, with digital footprints of varying sizes. Using xMIND, we systematically benchmark several content-based neural news recommenders (NNRs) in zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer scenarios, considering both monolingual and bilingual news consumption patterns. Our findings reveal that (i) current NNRs, even when based on a multilingual language model, suffer from substantial performance losses under ZS-XLT and that (ii) inclusion of target-language data in FS-XLT training has limited benefits, particularly when combined with a bilingual news consumption. Our findings thus warrant a broader research effort in multilingual and cross-lingual news recommendation. We release xMIND at https://github.com/andreeaiana/xMIND.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Multilingual and cross-lingual retrieval*; *Test collections.*

KEYWORDS

multilingual news dataset, news recommendation, low-resource languages, cross-lingual recommendation, machine translation

ACM Reference Format:
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 '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657867

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657867

1 INTRODUCTION

The digitalization of news consumption has established news platforms as the prevalent medium of information for Internet users. This, in turn, propelled personalized news recommendation systems into the main vehicle used by news websites to cater the individual information needs of readers. The global expansion of the Internet's outreach has vastly increased the language diversity of its users [41, 80], a non-negligible fraction of whom are polyglots, speaking and consuming news in two or more languages. For example, 22% of Americans speak a language other than English at home[1], whereas 65% of the working-age adults in the European Union know at least one foreign language.[2] However, the majority of online content remains available mostly in few resource-rich languages, with English accounting for more than half of the digital texts, while the majority of languages spoken around the globe appear in less than 0.1% of the websites or are not represented online.[3]

News media play a central role in democratic societies, by ensuring informed citizens and providing a public forum for disseminating and debating ideas and opinions [3, 22]. In this context, recommender systems shape people's worldviews and opinions through the way in which they filter and propagate news [45]. On the one hand, research on personalized neural news recommenders [70] has focused on improving their accuracy and diversity [56, 71, 73, 74], by mitigating technical challenges in news encoding [28, 40, 43, 49, 60, 65–68, 72, 78, 79] and user modeling [1, 26, 38, 50–52, 54, 61]. On the other hand, recent progress in areas such as neural machine translation (NMT) [10, 11, 15, 19, 35] and multilingual large language models (mPLMs) [8, 9, 58, 62–64, 77] for low(er)-resource languages has begun to democratize access to information for underrepresented communities [29].

Nonetheless, the existing body of research on news recommendation is limited in two main dimensions. Firstly, there is a *scarcity of publicly-available, diverse, multilingual news recommendation datasets* that could be leveraged to develop efficient multilingual news recommendation and support effective cross-lingual transfer to resource-lean languages. Despite the availability of adequate datasets being paramount to developing high-quality recommenders (e.g., see the Amazon dataset[4] for product recommendation, or MovieLens [21] and Netflix [5] for movie recommendation), the vast majority of news recommendation benchmarks are

[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

monolingual [13, 16, 17, 31, 44, 75]. Furthermore, the few existent multilingual benchmarks comprise only high-resource languages, and are either topic-specific and small-sized [23] or proprietary [72]. Secondly, the design of news recommenders for multi- and cross-lingual settings has been left largely unexplored. Traditional news recommendation systems support mostly monolingual recommendations, thus hindering simultaneous browsing for news across multiple languages. In practice, this translates into less relevant, less balanced and less diverse recommendations for multilingual news consumers [41]. Overall, these limitations pose significant challenges for online news readers who are multilingual or consume content in resource-lean and/or underrepresented languages.

Contributions. We address the above research gaps by introducing xMIND, a new public and large-scale multilingual news dataset for multi- and cross-lingual recommendation. xMIND is derived by translating the articles of the English-only dataset MIND [75] into a diverse selection of 14 high- and low-resource languages, spanning five geographical macro-areas and 13 distinct language families using the open-source machine translation system NLLB [10]. 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 areas, family languages, 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 TSV format and provide scripts for loading and combining the news with the corresponding click logs from the MIND dataset in the NewsRecLib [25] library.

We use xMIND to benchmark a range of state-of-the-art neural news recommenders, in zero-shot (`ZS-XLT`) and few-shot (`FS-XLT`) cross-lingual transfer setups, covering two realistic news consumption scenarios: monolingual and bilingual. We show that recommenders trained monolingually on English news suffer significant performance drops when evaluated on the target languages under `ZS-XLT`, 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, revealing the urgent need for developing more accurate and robust cross-lingual news recommendation approaches. Lastly, we investigate the quality of the translations in xMIND through a human-based annotation task and comparison against translations obtained with a commercial NMT system.

2 RELATED WORK

News Recommendation. Personalized news recommendation aims to alleviate the information overload of online news readers by providing suggestions tailored to their individual preferences [39, 70]. Neural news recommenders (NNRs) have become the driver of personalized news recommendation, replacing systems relying on manual feature engineering [70]. The majority of NNRs commonly consist of a dedicated (i) news encoder (NE), (ii) user encoder (UE), and a (iii) click predictor. The NE learns news representations from various input features (e.g., title, topical categories, named entities), either by instantiating convolutional neural networks [60, 65, 67], self-attention networks [50, 68], or graph attention networks [49] with pretrained word embeddings, or, more recently, by leveraging pretrained language models [40, 72, 78, 79]. Afterwards, the UE aggregates and contextualizes the embeddings of a user's clicked news into a user-level representation by means of sequential [1, 53, 61] or attentive [65, 68] encoders. Lastly, a candidate article's recommendation score is computed by comparing its embedding against the user profile [70]. Although a significant body of work has sought to improve personalization by enhancing NNRs' core components – news and user modeling – 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. [72] suggested instantiating the NE with mPLMs to enable news recommendation in diverse languages. Guo et al. [18] proposed a new NE 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.

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 [31] (German), Adressa [17] (Norwegian), Globo [13, 16] and its recent improved version NPR [44] (Portuguese), and MIND [75] (English). 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 [75]. However, these datasets consists only of monolingual news, and therefore, hinder the development of multilingual recommender systems. Iana et al. [23] aimed to address this problem by proposing NeMig, a multilingual news recommendation dataset in English and German. NeMig contains articles on the topic or 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. [72] mention the multilingual news recommendation dataset collected from the MSN News platform to analyze the effectiveness of mPLM-based NEs 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., it is not publicly available.

3 DATASET CREATION

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

Table 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 [10].

Code	Language	Script	Macro-area	Family	Genus	Total Speakers (M)	Res.	mPLM
SWH	Swahili	Latin	Africa	Niger-Congo	Bantu	71.6	high	yes
SOM	Somali	Latin	Africa	Afro-Asiatic	Lowland East Cushitic	22.0	low	yes
CMN	Mandarin Chinese	Han	Eurasia	Sino-Tibetan	Sinitic	1,138.2	high	yes
JPN	Japanese	Japanese	Eurasia	Japonic	Japanesic	1,234.5	high	yes
TUR	Turkish	Latin	Eurasia	Altaic	Turkic	90.0	high	yes
TAM	Tamil	Tamil	Eurasia	Dravidian	Dravidian	86.6	low	yes
VIE	Vietnamese	Latin	Eurasia	Austro-Asiatic	Vietic	85.8	high	yes
THA	Thai	Thai	Eurasia	Tai-Kadai	Kam-Tai	60.8	high	yes
RON	Romanian	Latin	Eurasia	Indo-European	Romance	24.5	high	yes
FIN	Finnish	Latin	Eurasia	Uralic	Finnic	5.6	high	yes
KAT	Georgian	Georgian	Eurasia	Kartvelian	Georgian-Zan	3.9	low	yes
HAT	Haitian Creole	Latin	North-America	Indo-European	Creoles and Pidgins	13.0	low	no
IND	Indonesian	Latin	Papunesia	Austronesian	Malayo-Sumbawan	199.1	high	yes
GRN	Guarani	Latin	South America	Tupian	Maweti-Guarani	(L1 only) 6.7	low	no

that is multi-parallel, i.e., where an article (i.e., a translation thereof) exists in each covered language. The former allows for a more realistic estimate of global multilingual and cross-lingual performance of news recommendation models [29, 47], 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 [75] (i.e., union of MINDlarge and MINDsmall) into 14 different languages using the NLLB 3.3B open-source NMT model [10]. 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, although 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 [24].

Language Selection. We select target languages for xMIND based on the following criteria: (1) linguistic diversity in terms of typological properties [14, 42], 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 [10]. Table 1 lists the selected languages, summarizing the following information, in accordance with the #BenderRule [4]:

- **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) [14]. We cross-reference the names with two other major linguistic resources, Glottolog [20] and Ethnologue [37];
- **Script**: We provide the English name of the script;
- **Family and Genus**: Language family and genus from WALS [14] and Glottolog [20];
- **Resource Level**: We borrow NLLB's [10] classification of languages into *low-* and *high-resource*;
- **mPLM Support**: We indicate whether the language is included in the pretraining corpora of XLM-RoBERTa [8], the representative mPLM used in our experiments (§4);
- **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]

[5]Note that 5.4% of the news in the entire MIND dataset do not contain an abstract.

Table 2: Indices of typological, genealogical, and geographical diversity for the language samples of different multilingual news recommendation datasets.

	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

We follow Ponti et al. [47] 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 [42]: each feature indicates the presence or absence of a particular linguistic property in a language. As per [47], 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 2 reports the three metrics for xMIND, as well as for NeMig [23] and the proprietary dataset from [72] (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,

[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 [14], are: Africa, Australia, Eurasia, North America, Papunesia, and South America.

Table 3: Statistics of the subset of Global Voices [57] used as validation data for tuning the NLLB [10] hyperparameters.

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 4: Hyperparameter optimization results on the subset of Global Voices [57]. We report only the results obtained with the best number of beams. We report macro-average sacreBLEU scores over six language pairs.

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

and 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 [10]. 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 [8]. 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 [10] using a subset of Global Voices (GV) [57] 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 of xMIND as the target: this results in six language pairs, statistics of which are reported in Table 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 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 [48]. 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 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/

Table 5: Number of news in the different splits of xMIND.

Small		Large		
Train	Dev	Train	Dev	Test
51,282	42,416	101,527	72,023	120,959

Final Dataset. The xMIND dataset contains 130,379 unique news in the 14 different languages listed in Table 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 [75]. Following Wu et al. [75], 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.[11] 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 [75], using the news IDs. Additionally, to facilitate a seamless integration with existing NNRs, we implement the data loading functionality for xMIND in NewsRecLib [25].

4 EXPERIMENTAL SETUP

We systematically benchmark a range of state-of-the-art content-based NNRs in zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer scenarios. Our experiments encompass two types of news consumption patterns: monolingual and bilingual.

4.1 Benchmarked Recommenders

Neural News Recommenders. We evaluate several content-based NNRs: (1) *NAML* [65], (2) *LSTUR* [1], (3) *MINS* [61], (4) *CAUM* [52], (5) *TANR* [67], (5) *MINER* [38], and (6) *MANNeR* [27] (only the CR-Module responsible for pure content-based recommendation, without any aspect-based personalization or diversification). Additionally, we use as baseline (7) $NAML_{CAT}$, a language-agnostic variant of NAML which learns news embeddings solely based on randomly-initialized category vectors, and user representations by attending over the embeddings of the clicked news. With the exception of MINER and MANNeR, designed with a PLM-based NE, the remaining NNRs originally contextualize word embeddings with convolutional neural networks (CNNs) [32], additive attention [2], or multi-head self-attention [59] networks. For fair comparison and to enable multilingual recommendations, we follow Wu et al. [72], and replace the original NEs 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 differ in their UE component: NAML [65] and TANR [67] encode users' preferences using additive attention; MINS [61] combines multi-head self-attention with a multi-channel GRU-based [7] recurrent network and additive attention;

[11]https://msnews.github.io/

MINER [38] introduces a poly-attention approach based on multiple additive attentions to learn various interest vectors for each user. Moreover, 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 [1].[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 [52]. In contrast to the other models, MANNeR does not learn a parameterized UE, instead using a late fusion approach consisting of the mean-pooling of dot-product scores between each of the candidates and the clicked news [27].

All benchmarked models compute the recommendation score as the dot product between the representations of the candidate and the clicked news. MANNeR is optimized using a supervised contrastive loss (SCL) [30], whereas the other models are trained by minimizing the standard cross-entropy loss.

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. Since Wu et al. [75] do not release test labels for MIND, we use the validation portion for testing, and split the respective training set into temporally disjoint training (first four days) and validation (last day) sets.

Training Details. We use XLM-RoBERTa [8] as the mPLM in all models, and fine-tune only its last four layers.[13] We use 100-dimensional TransE embeddings [6], pretrained on Wikidata, to initialize the entity encoder in the NE 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. [69]. To ensure comparability, we train all models with mixed precision using the NewsRecLib[15] library [25], with a batch size of 8, for 10 epochs with early stopping, optimizing with the Adam algorithm [33].

We perform hyperparameter optimization for the most important hyperparameters of each NNR using 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. We optimize the number of heads in the multi-head self-attention networks of NAML, LSTUR, MINS, TANR, and CAUM in $[8, 12, 16, 24, 32]$, and the query vector dimensionality in the additive attention network in $[50, 200]$ with a step of 50 for NAML, LSTUR, MINS, and TANR. Moreover, we search the optimal SCL temperature in MANNeR sweeping the interval $[0.1, 0.5]$, with a step of 0.02. Lastly, for MINER, we optimize the number of context codes in the interval $[8, 16, 32, 48]$, and choose the best-performing aggregation type between *mean*, *max*, and *weighted*. 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, and report averages and standard deviations for the standard metrics: AUC, MRR, nDCG@5, and nDCG@10.[16]

4.2 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 on English (i.e., on the MIND news) to generate recommendations in another language (i.e., in one of the 14 languages of 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 increasingly replace a portion (varying from 10% to 90%) of the English training set (both in history and candidate set) with target-language news. For a fair setup (i.e., no knowledge of test data distributions during training), the distribution of languages in our validation sets mirror the language ratios of respective training sets [55].

We couple the two training scenarios (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 randomly replace a portion of the English news with corresponding xMIND translations in the target language. Like in bilingual training, we also vary the portion of replaced news in the interval $[10\%, 90\%]$, with a 10% step.

The two training setups, each combined with both consumption patterns, thus result in four types of experiments: **(i) ZS-XLT$_{\text{MONO}}$** – monolingual training (in English) and evaluation on monolingual news consumption in the target language; **(ii) ZS-XLT$_{\text{BILING}}$** – monolingual training (in English) and evaluation on bilingual news consumption in English and the target language; **(iii) FS-XLT$_{\text{MONO}}$** – bilingual training in a mixture of English and target-language and evaluation on monolingual news consumption in the target language **(iv) FS-XLT$_{\text{BILING}}$** – bilingual training in a mixture of English and target-language and evaluation on bilingual news consumption in English and the target language.[17]

5 RESULTS AND DISCUSSION

We first analyze the ZS-XLT performance of the benchmarked models, comparing it with their performance on the English data (i.e., MIND). Then we examine the NNRs' capabilities in FS-XLT, to determine the extent to which having (some) training data in the target language influences the recommendation performance. An

[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 [1] for more details.

[13] In the interest of computational efficiency, we keep the bottom eight layers of the Transformer encoder frozen.

[14] The entity vectors are provided as part of the original MIND dataset [75].

[15] https://github.com/andreeaiana/newsreclib

[16] We train all models on one NVIDIA Tesla A100 (with 40/80 GB memory) or A40 (with 48 GB memory).

[17] Note that due to the high computational costs, in the case of FS-XLT$_{\text{BILING}}$ 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).

Table 6: ZS-XLT$_{\text{MONO}}$ recommendation performance. For each model, we report 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, standard deviations across three runs. The best results per column are highlighted in bold, the second best are underlined.

Model	AUC			MRR			nDCG@5			nDCG@10		
	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ
NAML$_{\text{CAT}}$	55.46±0.18		0.0	31.12±0.56		0.0	29.44±0.67		0.0	35.81±0.59		0.0
CAUM-PLM	57.82±3.01	55.90±1.75	-3.32	32.92±1.68	31.38±1.62	-4.68	31.09±1.88	29.60±1.76	-4.77	37.49±1.71	35.96±1.58	-4.08
LSTUR-PLM	56.80±1.36	56.28±1.68	-0.92	33.00±0.59	31.53±0.85	**-4.47**	31.18±0.54	29.70±0.92	**-4.73**	37.45±0.54	36.03±0.85	**-3.78**
MANNeR	50.00±0.00	50.00±0.00	**0.00**	35.58±0.31	33.03±0.54	-7.15	33.86±0.22	31.34±0.48	-7.45	40.17±0.21	37.64±0.44	-6.28
MINER	57.73±7.33	55.81±4.33	-3.32	31.71±4.95	30.20±3.42	-4.76	30.01±4.95	28.53±3.52	-4.91	36.45±4.84	35.02±3.51	-3.90
MINS-PLM	**59.89±0.29**	**56.94±1.40**	-4.93	34.75±0.24	33.11±0.51	-4.70	32.94±0.23	31.32±0.52	-4.93	39.35±0.20	37.64±0.50	-4.35
NAML-PLM	52.85±2.27	52.49±2.60	-0.68	**35.98±0.44**	**33.98±0.95**	-5.56	**34.11±0.46**	**32.13±1.08**	-5.80	**40.43±0.39**	**38.38±1.02**	-5.06
TANR-PLM	54.18±5.91	53.27±1.91	-1.68	35.47±0.95	32.14±0.9	-9.40	33.56±1.07	30.25±1.02	-9.98	40.03±0.86	36.78±0.88	-8.11

ideal NNR should rank positive candidates higher than negative ones regardless of their language. We thus focus our analysis mostly on the ranking performance (i.e., the nDCG@10 metric).

5.1 Zero-Shot Cross-Lingual Transfer

Table 6 shows the ZS-XLT$_{\text{MONO}}$ recommendation performance of the models, averaged over the 14 languages of xMIND, in comparison to their recommendation results on MIND (i.e, when both trained and evaluated in English). To contextualize these findings, we additionally report the performance of NAML$_{\text{CAT}}$, which embeds news based only on their topical categories, completely ignoring the content. As a content/language-agnostic recommender, NAML$_{\text{CAT}}$ is thus an apt baseline both in English and the target languages.

Firstly, in English recommendation on MIND, all NNRs improve over the category-based recommender, with relative improvements in nDCG@10 ranging from 1.78% (i.e., MINER) to 12.91% (i.e., NAML-PLM). In ZS-XLT$_{\text{MONO}}$, however, NNRs exhibit weaker performance w.r.t. the content-agnostic NAML$_{\text{CAT}}$, with at most 7.19% relative gain (i.e., NAML-PLM), and in some cases even lower performance (e.g., MINER with 2.19% relative drop). Secondly, we find that NNRs trained monolingually in English and evaluated on the target languages of xMIND suffer an average relative drop in performance between 3.78% (i.e., LSTUR-PLM) and 8.11% (i.e., TANR-PLM) w.r.t. to their English performance. Models that achieve the best (ranking) performance in English (e.g., NAML-PLM, MANNeR, MINS) still (i) perform best in ZS-XLT$_{\text{MONO}}$, exhibiting, however, (ii) the highest relative performance drops w.r.t. their English performance. While this might lead to the conclusion that such NNRs are less robust in XLT, one should keep in mind that a *random* recommender would exhibit the same performance regardless of the language of the news (i.e., 0% drop in XLT performance w.r.t. English). In other words, the absolute XLT performance of NNRs still matters more than the relative drops w.r.t. their English performance.

We next analyze the models' ZS-XLT$_{\text{MONO}}$ performance across the 14 target languages individually, captured by Fig. 1. As expected, the NNRs achieve the best results on VIE, IND, RON, and FIN, which constitute some of the most represented languages in terms of numbers of tokens in the pretraining corpora of XLM-RoBERTa [8]. At the same time, we observe the lowest performance across models for KAT, HAT, and GRN, with the latter two languages being out-of-sample (i.e., not seen in pretraining) for XLM-RoBERTa.

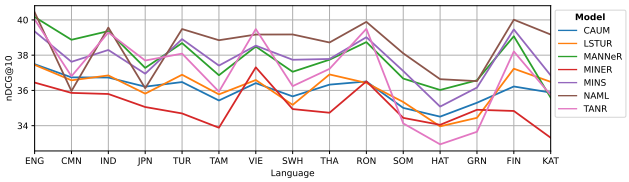


Figure 1: ZS-XLT$_{\text{MONO}}$ ranking performance, w.r.t. nDCG@10, across the 14 languages in xMIND and English.

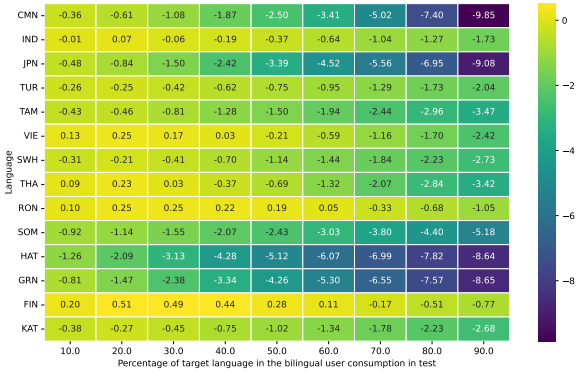
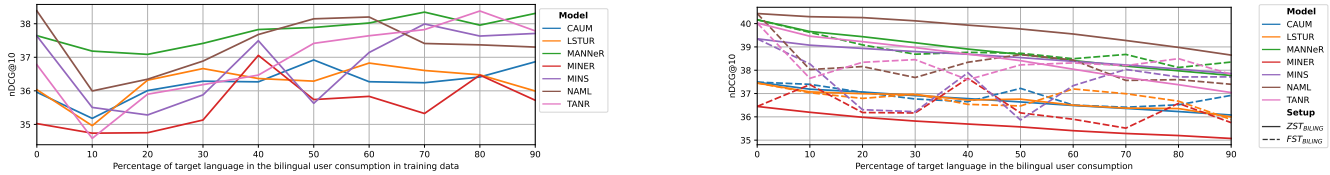


Figure 2: Relative percentage difference in ranking performance (w.r.t. nDCG@10), under ZS-XLT$_{\text{BILING}}$ compared to full English training and testing, for NAML-PLM.

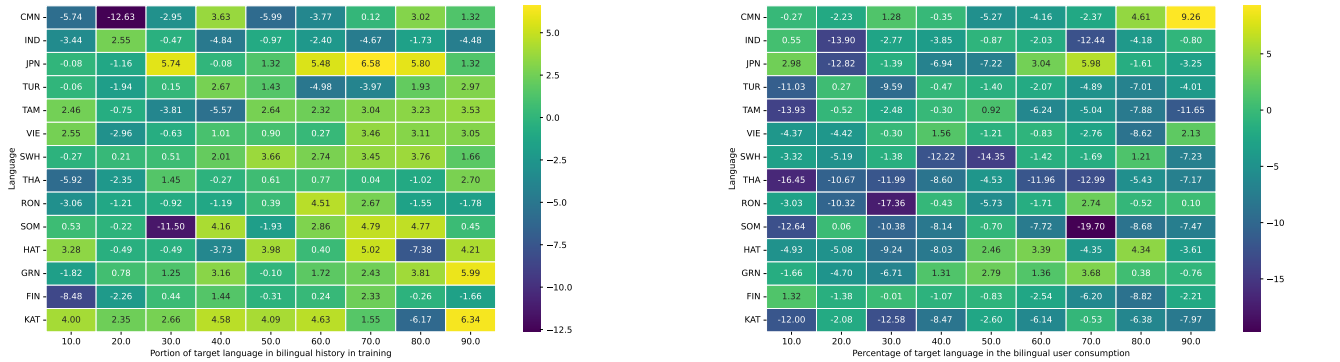
Next, we examine the change in the models' ranking performance when the user consumption is bilingual (i.e., ZS-XLT$_{\text{BILING}}$). Fig. 2 shows the results relative to the corresponding English performance for NAML-PLM (i) across target languages and (ii) for varying percentages of the target language in the user's news consumption. Overall, we notice a steady decrease in performance for all models correlated with higher percentages of target language in the consumption pattern. Additionally, the performance of all NNRs deteriorates most drastically for the languages unseen by the mPLM during pretraining, namely HAT and GRN. Surprisingly, although based on the same mPLM, the NNRs are not equally robust to the choice of the target language. For example, NAML-PLM's



(a) FS-XLT$_{MONO}$ setup.

(b) FS-XLT$_{BILING}$ setup.

Figure 3: FS-XLT ranking performance, averaged over the 14 languages of xMIND, for various portions of target language in the user's bilingual news consumption.



(a) FS-XLT$_{MONO}$ compared to ZS-XLT$_{MONO}$ ranking performance for MANNeR.

(b) FS-XLT$_{BILING}$ compared to ZS-XLT$_{BILING}$ ranking performance for NAML-PLM.

Figure 4: Relative percentage difference in ranking performance (w.r.t. nDCG@10), under FS-XLT compared to ZS-XLT.

performance drops up to 9.85% when the user history contains CMN – a higher decrease than for HAT or GRN – whereas MANNeR's (relative) drop is at only 3.16%. Additionally, we find that for some languages and models (e.g., LSTUR-PLM and RON or FIN), the decline in performance is lower when either English or the target language represent the predominant language in the user's news consumption. Given the models' diverse designs, this points to the need to investigate the robustness of the architecture, particularly of the UE, to changes in the composition of the user history, which so far, has only been assumed to be monolingual.

5.2 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 [55]. 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 [36]. Motivated by these findings, we further analyze whether adding target-language data in training also benefits news recommenders: Fig. 3a shows the NNRs' ranking performance (nDCG@10), averaged over all 14 target languages, for increasing percentages of target-language news included in the training data. Our results show that incorporating some target language data in training indeed ameliorates the performance losses from ZS-XLT$_{MONO}$. However, we find that if the target language constitutes less than 30-40% of training news the recommendation performance drops (compared to ZS-XLT). We believe that this is due to the relatively short user histories, consisting of 33 articles on average: a small percentage of news in another language is likely to confound the recommenders' UE. Moreover, although higher portions of target-language training data lead, on average, to higher gains over ZS-XLT, the gains vary per model and percentage of target-language news injected. More specifically, models such as NAML-PLM or CAUM-PLM tend to produce less accurate rankings when the two languages seen during training are unevenly represented (e.g., for low and high portions of the target language). Delving deeper into per-language results (Fig. 4a), we find that FS-XLT particularly benefits low-resource languages and languages unseen in pretraining of XLM-RoBERTa, as the mPLM on which the NE of all NNRs in our evaluation are based.

In contrast to the FS-XLT$_{MONO}$ setting, few-shot target-language injection does not appear to be equally effective when we assume a bilingual news consumption of users during evaluation, that is, in the FS-XLT$_{BILING}$ scenario. Fig. 3b compares the ranking performance (nDCG@10) of NNRs in FS-XLT$_{BILING}$ against their respective performance in ZS-XLT$_{BILING}$, for varying proportions of target-language news included in the training. The majority of NNRs perform on par or better when a few instances in the target languages are seen during training. However, the performance of NAML-PLM – generally the best performing NNR in our evaluation – is subpar to that achieved by the model trained only on English. A closer look at its performance across languages, shown in Fig 4b, reveals that FS-XLT benefits the model primarily for those languages for which it exhibits the highest losses under ZS-XLT$_{BILING}$,

	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 5: Annotator agreement in terms of Krippendorf's alpha, per language, for all questions in the annotation task.

namely CMN, JPN, HAT, and GRN (i.e., Fig. 2). For other NNRs, the gains are more evenly distributed across all languages.

Overall, these variations in the performance of models and the limited benefits of few-shot target-language injection – particularly when considering bilingual news consumption – emphasize again the need for a deeper understanding of the specific factors that drive multilingual NNR performance in order to inform design of user encoder architectures that are robust to multilingual user histories.

6 TRANSLATION QUALITY

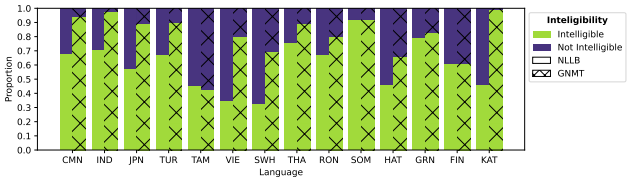
We finally investigate the quality of the translations in xMIND. Concretely, we (i) estimate the translation quality and (ii) investigate the robustness of the NNRs to different translations of the source news from MIND [75]. To this end, we use xMINDsmall and additionally translate its training and development portions using Google (Neural Machine) Translation (GNMT) [76], a commercial MT system that supports all xMIND languages. GNMT has been shown to outperform NLLB on translation from English to various low-resource languages [10]. [18] [19]

6.1 Manual Quality Estimation of Translation

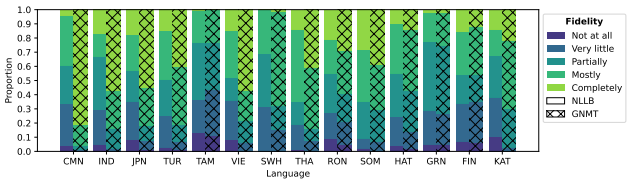
Given the size of the xMINDsmall dataset and our (limited) annotation budget, it was infeasible to manually post-edit the translations of the news in the test portion of xMIND. We therefore resorted to conducting an annotation task to estimate the quality of the translations. To this end, we sample 50 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 [46]. Two annotators judged the quality of the NLLB and GNMT translations for each language.[20] 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

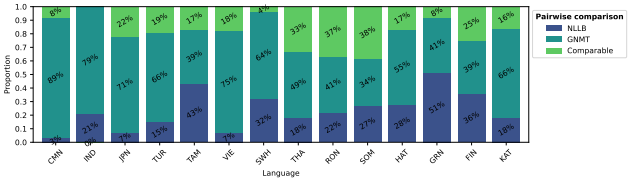
[18] GNMT is a proprietary system, hindering fair comparisons to open-source models due to the lack of transparency regarding its model architecture and training procedures.
[19] We translated the text with the Cloud Translation - Advanced (v3) API: https://cloud.google.com/translate/docs/overview, using Google Cloud research credits worth approximately $5,000.
[20] 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.



(a) Intelligibility.



(b) Fidelity.



(c) Pairwise comparison.

Figure 6: Annotation task results for each question type.

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 [34]) for the NLLB translations for the first two questions, as shown in Fig. 5, 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. 6a). 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. 6b. Notably, the NLLB-sourced translation 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, shown in Fig. 6c, which show that NLLB translations are judged to be overall better than their GNMT counterparts for these two 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)

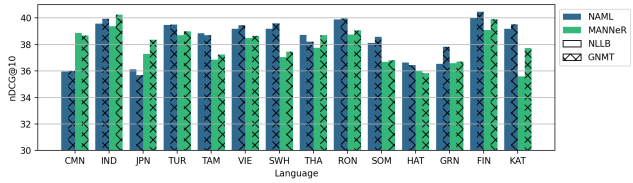
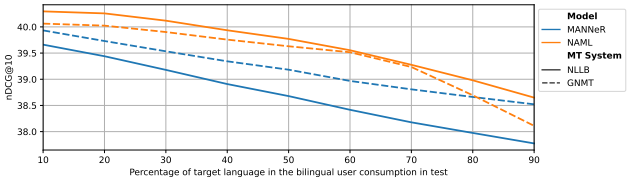
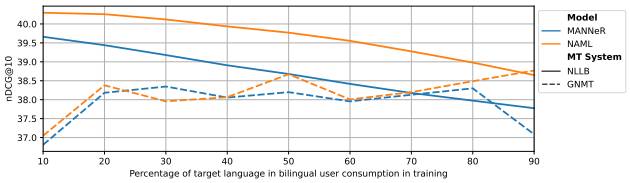


Figure 7: ZS-XLT$_{MONO}$ ranking performance, w.r.t. nDCG@10, in terms of MT system, for NAML-PLM and MANNeR.



(a) ZS-XLT$_{BILING}$ ranking performance.



(b) FS-XLT$_{MONO}$ ranking performance.

Figure 8: Ranking performance, w.r.t. nDCG@10, in terms of MT system, for NAML-PLM and MANNeR.

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 [75], 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 [12]. 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 [10].

6.2 Robustness of NNRs to Translation Quality

We next investigate the robustness of the best-performing NNRs from our previous experiments, namely NAML-PLM 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. 7 shows the ranking performance (w.r.t. nDCG@10) for both models and MT systems under ZS-XLT$_{MONO}$ (i.e., training on the English MINDsmall 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$_{BILING}$ setting. The corresponding results from Fig. 8a 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 8b illustrates the ranking performance for the FS-XLT$_{MONO}$ 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.

7 CONCLUSION

The ever-growing language-diversity of online news readers has not been reflected in the news recommendation research, which focuses nearly entirely on resource-rich languages, particularly English, and monolingual news consumption. In this work, we introduced xMIND, an open multilingual news recommendation dataset comprising 14 linguistically and geographically diverse languages, derived from the English MIND dataset using machine translation. We used xMIND to benchmark several state-of-the-art content-based NNRs in both zero-shot and few-shot cross-lingual recommendation transfer, experimenting with both monolingual and bilingual news consumption patterns. Our findings show that current NNRs suffer considerable performance losses under ZS-XLT, while the inclusion of target-language data in FS-XLT training brings limited gains to recommenders, especially in the context of bilingual news consumption. We believe that xMIND is a valuable resource for the news recommendation community, and hope it will foster much more research on multilingual and cross-lingual news recommendation, for speakers of both high- and low-resource languages.

ACKNOWLEDGMENTS

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. This material is based upon work supported by the Google Cloud Research Credits program with the award EDU275608761.

REFERENCES

[1] 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*. 336–345.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ICLR* (2014).

[3] Jack M Balkin. 2017. Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UCDL rev.* 51 (2017), 1149.

[4] Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient* 14 (2019).

[5] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. New York, 35.

[6] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. 2787–2795. https://dl.acm.org/doi/abs/10.5555/2999792.2999923

[7] Kyunghyun Cho, Bart van Merriënboer, Çağlar 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)*. 1724–1734. https://doi.org/10.3115/v1/D14-1179

[8] 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*. 8440–8451.

[9] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 7059–7069.

[10] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

[11] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.

[12] 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)*. 3608–3626.

[13] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. 2018. News session-based recommendations using deep neural networks. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*. 15–23.

[14] Matthew S. Dryer and Martin Haspelmath (Eds.). 2013. *WALS Online (v2020.3)*. Zenodo. https://doi.org/10.5281/zenodo.7385533

[15] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research* 22, 107 (2021), 1–48.

[16] 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 (2019), 169185–169203.

[17] 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*. 1042–1048.

[18] 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*. 1130–1140.

[19] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics* 48, 3 (2022), 673–732.

[20] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. glottolog/glottolog: Glottolog database 4.4.

[21] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[22] Natali Helberger. 2021. On the democratic role of news recommenders. In *Algorithms, Automation, and News*. Routledge, 14–33.

[23] Andreea Iana, Mehwish Alam, Alexander Grote, Katharina Luwig, Philipp Müller, Christof Weinhardt, and Heiko Paulheim. 2023. NeMig-A Bilingual News Collection and Knowledge Graph about Migration. In *Proceedings of the Workshop on News Recommendation and Analytics co-located with RecSys 2023*.

[24] Andreea Iana, Mehwish Alam, and Heiko Paulheim. [n. d.]. A survey on knowledge-aware news recommender systems. *Semantic Web* Preprint ([n. d.]), 1–62.

[25] 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*. 296–310.

[26] 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*. 2384–2388.

[27] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023. Train once, use flexibly: A modular framework for multi-aspect neural news recommendation. *arXiv preprint arXiv:2307.16089* (2023).

[28] 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*. 15647–15658.

[29] 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*. 6282–6293.

[30] 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 (2020), 18661–18673.

[31] 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*. 16–23.

[32] 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, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[33] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR* (2014).

[34] Klaus Krippendorff. 2013. *Content analysis: An introduction to its methodology*. Sage publications.

[35] 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*.

[36] 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)*. 4483–4499.

[37] 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 (2009), 2010.

[38] 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*. 343–352.

[39] Miaomiao Li and Licheng Wang. 2019. A survey on personalized news recommendation technology. *IEEE Access* 7 (2019), 145861–145879.

[40] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. PBNR: Prompt-based News Recommender System. *arXiv preprint arXiv:2304.07862* (2023).

[41] 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 (2020), 516–535.

[42] 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*. 8–14.

[43] Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. 2020. KRED: Knowledge-aware document representation for news recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 200–209. https://doi.org/10.1145/3383313.3412237

[44] 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)*.

[45] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[46] 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*. 327–337.

[47] 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)*. 2362–2376.

[48] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (Eds.). Association for Computational Linguistics, Brussels, Belgium, 186–191. https://doi.org/10.18653/v1/W18-6319

[49] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. Personalized news recommendation with knowledge-aware interactive matching. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 61–70. https://doi.org/10.1145/3404835.3462861

[50] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. 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)*. 5457–5467. https://doi.org/10.18653/v1/2021.acl-long.424

[51] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. 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*. 1974–1978.

[52] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News recommendation with candidate-aware user modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1917–1921.

[53] 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*. 1423–1432. https://doi.org/10.18653/v1/2020.findings-emnlp.128

[54] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. 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)*. 5446–5456. https://doi.org/10.18653/v1/2021.acl-long.423

[55] 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*. 10725–10742.

[56] Hao Shi, Zi-Jiao Wang, and Lan-Ru Zhai. 2022. DCAN: Diversified news recommendation with coverage-attentive networks. *arXiv preprint arXiv:2206.02627* (2022). https://doi.org/10.48550/arXiv.2206.02627

[57] Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010. https://dl.acm.org/doi/abs/10.5555/3295222.3295349

[60] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*. 1835–1844. https://doi.org/10.1145/3178876.3186175

[61] Rongyao Wang, Shoujin Wang, Wenpeng Lu, and Xueping Peng. 2022. News recommendation via multi-interest news sequence modelling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7942–7946.

[62] 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* (2023).

[63] 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*.

[64] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[65] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3863–3869. https://doi.org/10.24963/ijcai.2019/536

[66] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2576–2584. https://doi.org/10.1145/3292500.3330665

[67] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual meeting of the association for computational linguistics*. 1154–1159. https://doi.org/10.18653/v1/P19-1110

[68] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. 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)*. 6389–6394. https://doi.org/10.18653/v1/D19-1671

[69] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2022. Rethinking InfoNCE: How Many Negative Samples Do You Need?. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 2509–2515. https://doi.org/10.24963/ijcai.2022/348

[70] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems* 41, 1 (2023), 1–50.

[71] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. 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*. 44–53. https://aclanthology.org/2020.aacl-main.6

[72] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[73] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. End-to-end Learnable Diversity-aware News Recommendation. *arXiv preprint arXiv:2204.00539* (2022). https://doi.org/10.48550/arXiv.2204.00539

[74] Chuhan Wu, Fangzhao Wu, Tao Qi, Wei-Qiang Zhang, Xing Xie, and Yongfeng Huang. 2022. Removing AI's sentiment manipulation of personalized news delivery. *Humanities and Social Sciences Communications* 9, 1 (2022), 1–9.

[75] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3597–3606.

[76] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[77] 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*. 483–498.

[78] 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*. 5478–5489. https://aclanthology.org/2022.emnlp-main.368

[79] Zizhuo Zhang and Bang Wang. 2023. Prompt learning for news recommendation. *arXiv preprint arXiv:2304.05263* (2023).

[80] Ethan Zuckerman. 2008. The polyglot internet. (2008). https://ethanzuckerman.com/the-polyglot-internet/