LLM4DDC: Adopting Large Language Models for Research Data Classification Using Dewey Decimal Classification

Gautam Kishore Shahi [1], Renat Shigapov [2], Oliver Hummel [1]

[1] Technical University of Applied Sciences, Mannheim, Germany;
[2] University of Mannheim, Mannheim, Germany

Classifying research data in institutional repositories is time-consuming and challenging. While the Dewey Decimal Classification (DDC) system is widely used in subject classification for texts, its application to research data metadata has been limited so far. This study explores the possible use of large language models (LLMs) and small language models (SLMs) for the automatic classification of research data in the context of DDC. This study uses sample data from an existing dataset compiled from different institutions mainly in Germany. We use a prompt engineering approach for LLMs, and fine tuning for SLMs, where we use RoBERTa as a baseline. Our results show that LLMs with prompt engineering currently are not able to classify metadata of research data into DDC classes as good as SLMs with fine tuning. To foster adoption, we openly release our models, code, and datasets for integration into research data infrastructures at GitHub.[1]

Keywords: Dewey Decimal Classification, Research Data, Large Language Model, Automated Classification

1 Introduction

As the volume of research data continues to grow, accurately classifying this data in institutional repositories remains a significant challenge, as it requires significant human

1 https://github.com/TransforMA-WP3/LLM4DDC; *Visited by the editors on June 30, 2025.*

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23948 (CC BY-SA 4.0).

effort to assign the appropriate subject classification. While the Dewey Decimal Classification (DDC) system is widely used for automatic subject indexing (Golub 2021) in the context of libraries, its application to automating the creation of metadata for research data is at best a fledgling discipline (Weber et al. 2020). We address this gap by exploring the use of large language models (LLMs) and small language models (SLMs) in automating the detection of research areas for a DDC classification of research data (Hu et al. 2024). This has practical implications for numerous data repositories, including the German National Research Data Infrastructure (NFDI), where accurate metadata is crucial for effective research data management and retrieval.

Different approaches have been tested for the automatic classification of research data. Recent research (Ho et al. 2024) used traditional Machine Learning (ML) models for identifying DDC classifications, having DDC classes in the range of 400 to 499. Our current focus is on identifying DDC classes at the domain and subject level out of the three-level DDC classification system (ref. Section 2.2) to better understand the feasibility of such an approach. For the LLMs we were aiming on optimizing model parameters such as temperature and leveraging prompt-engineering strategies with zero-shot and few-shot prompts (Shahi and Hummel 2025). For SLMs, we used fine-tuned versions of pre-trained models, such as RoBERTa from Hugging Face[2]. In both settings, we used title and description of the research data to identify the best matching DDC classes and evaluated the results with precision, recall and F1-score.

The remainder of the paper is organized as follows; Section 2 describes the background, Section 3 describes an experiment, Section 4 describes the results and their analysis, before finally, Section 5 provides the conclusion and discusses future work.

2 Background

2.1 Contemporary Language Models

Large Language Models (LLMs) have evolved significantly over the last three years based on the advancement of deep learning which was mainly caused by improved computing performance through GPUs as well as the availability of large-scale datasets for model training (Zhou et al. 2024). Early language models, such as word embeddings (Word2Vec Church 2017), laid the foundation for contextualized and semantic text representations. The introduction of transformer-based architectures, such as Google's bidirectional transformers for language understanding (BERT; Devlin et al. 2019), revolutionized natural language processing (NLP) by enabling models to generate coherent and contextually rich text. However, BERT is still considered being a Small Language Model (SLM)[3].

2 https://huggingface.co/; *Visited by the editors on June 30, 2025.*

3 https://www.ibm.com/think/topics/small-language-models; *Visited by the editors on June 30, 2025.*

In scholarly communication, LLMs have been increasingly utilized to enhance various aspects of scientific publishing. Researchers have also been leveraging LLMs for subject tagging (Shahi and Hummel 2025) or extracting metadata from papers (Watanabe, Ito, and Matsubara 2025). Moreover, SLMs have already been used for assigning DDC classes to scientific papers (Ho et al. 2024). Hence, it seems worthwhile to explore both SLMs and LLMs for DDC classification in this study.

2.2 DDC

The Dewey Decimal Classification (DDC) system is a method for organizing books and other materials in libraries. Created by Melvil Dewey in 1876 (Dewey 1899), it assigns a unique number to each subject, making it easy to group publications and find related materials. We used the 23rd edition of DDC released in 2011 (Dewey et al. 2011). The classification system is divided into three levels; the first level consists of ten domains, then each domain has ten subjects, each comprising ten more specific topics; hence the "decimal" in the name. Each domain, subject, and topic is assigned a number (Dewey 1899). For example, 500 is assigned for domain *science*, while 530 stands for *Physics* and 532 is *Fluid Mechanics* as subcategory of science and physics. Due to its simplicity, in combination with a comparatively good topic coverage, the DDC is widely used in libraries worldwide (Wang 2009).

3 Approach & Experiments

The approach we used to classify research data into their DDC classes is illustrated in Figure 1. The pipeline consists of dataset collection, data cleaning, building classification models based on contemporary LLMs and SLMs, and finally evaluating the results obtained. Each step is explained in more detail below. The replication package for this work is available at (Shahi, Shigapov, and Hummel 2025).

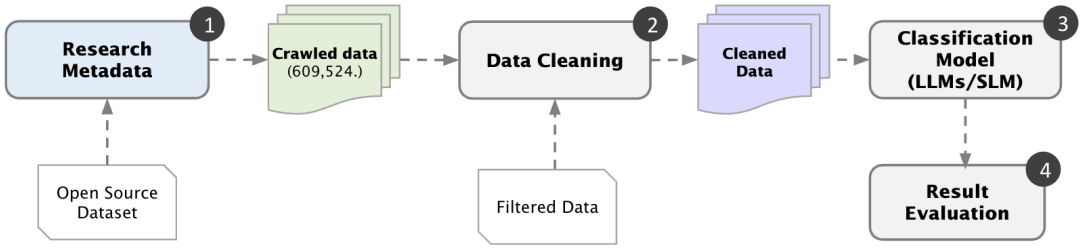


Figure 1: Overview of evaluation approach.

3.1 Dataset

We used a list of research data collected from Weber et al. (2020) as ground truth. This dataset is compiled from 29 different sources such as various German university libraries, Arxiv and some other data sources. In total, Weber et al. have collected around 16 million records, however, the final dataset they have shared consists of only roughly 610,000 records. Out of this, we filtered around 222,000 records actually having valid DDC classes.

Moreover, the dataset was highly imbalanced, so we randomly sampled 1,000 entries from those 14 subjects actually containing more than 1,000 entries, obtaining the total of 14,000 records used in this study. The subjects spanned 6 different domains (Science, Technology, Computer Science Information & General Works, Language, Philosophy & Psychology, Social Sciences).

Preprocessing & Cleaning

Prior to selecting the actual records for our study, we cleaned the dataset by removing research data titles or texts that were not in English or had an obviously wrong DDC number assigned, such as *488348*. Moreover, we filtered out records that included only a broad domain without specifying a subject. For example, DDC number 551 corresponds to the domain *Science*, but no specific subject is provided, which would have caused issues in our experiments.

3.2 Classification Model

We used the cleaned data for the evaluation of our classification models. A concatenation of text and title was used to yield DDC number and class name from the LLMs and SLMs. Both classification models were implemented for the domain and subject levels of DDC (cf. Section 3.1).

Evaluated Models

We used the following freely available open-source LLMs based on the Ollama framework for our experiments.[4] Ollama is a tool designed to run LLMs containing a large variety of LLMs. It allows downloading LLMs with various parameter sizes. We chose to use Llama,

4 https://ollama.com/library; *Visited by the editors on June 30, 2025.*

Mistral, and Gemma as described in the following: *Llama 3.1* is a state-of-the-art open-source language model developed by Meta for natural language processing tasks. In our experiments, we used the 70-billion-parameter version. *Gemma* is a family of lightweight language models created by Google DeepMind for efficient NLP. We utilized Gemma 2.7, which has 27 billion parameters. *MistralLite* is a compact and efficient language model developed by Mistral AI, optimized for fast inference and strong performance across a range of NLP tasks. It is the smallest model in our setup, with only 7 billion parameters.

To invoke the above LLMs, we used zero and few-shot prompts as has proven successful in our previous work (Shahi and Hummel 2025). In the zero-shot (prompt 1), we ask the LLM to classify research data for DDC without providing any additional explanation. In the few-shot prompt 2, we provided the hierarchical structure of DDC domain and subjects and asked the LLM for an appropriate classification. In both prompts, we asked for DDC numbers and classes, which we then used for evaluation (cf. Section 4).

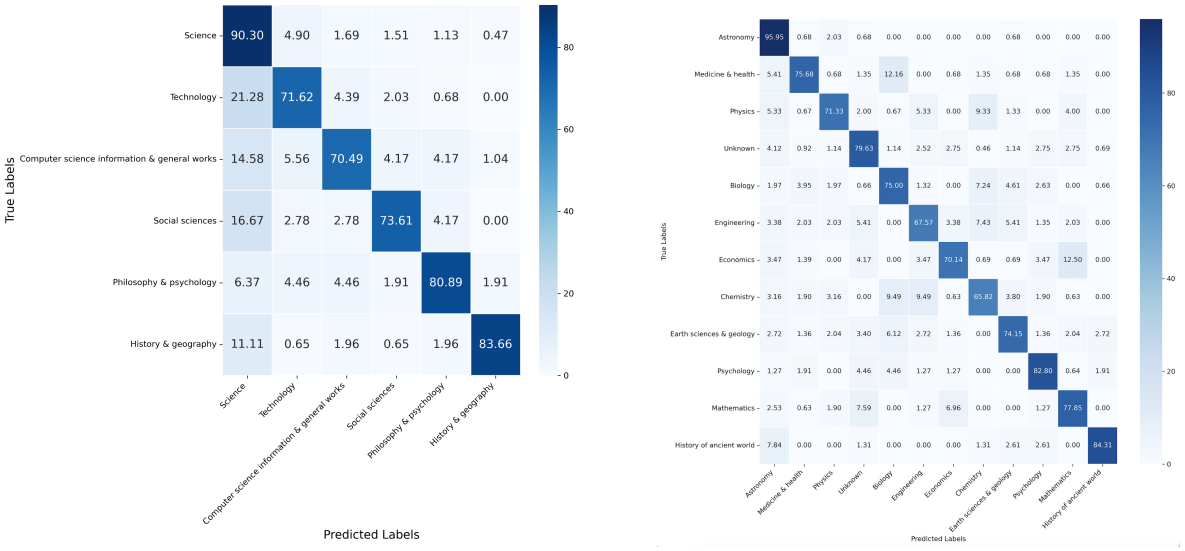
For the baseline comparison with an SLM, we used the pre-trained baseline models *BERT* (Devlin et al. 2019), *RoBERTa* (Liu et al. 2019), and *DistilBERT* (Sanh et al. 2019) model from Huggingface and fine-tuned them with with a part of the dataset described above. For these classification models, text was used as feature and domain and subject names were provided as target variable.

4 Evaluation & Result Analysis

We evaluated the classification performance with precision, recall, and F1 scores, calculated separately for both classification levels: domain and subject. At the domain level, we considered six different domains, while the subject level test data included 14 distinct classes overall.

For the LLMs, the F1 score ranged from 0.15 to 0.43, and they struggled to predict both DDC numbers and subjects accurately. In contrast, the fine-tuned SLMs performed much better, as we found that RoBERTa performed best from all three models (BERT, RoBERTa, DistilBERT), achieving F1 scores between 0.83 and 0.96 (in the subject Astronomy 520). Figure 2 shows the confusion matrix for the best domain and subject classification yielded with RoBERTa.

A manual analysis of the LLM's outputs revealed that they often provided only partial DDC classifications. For instance, Llama (LLM) predicted Philosophy instead of the correct class Philosophy & Psychology, suggesting a potential leverage for relatively simple future improvements. Overall, classification at the subject level was more accurate than at the domain level. The best results were obtained for the Science domain and its subjects, followed by Technology.



(a) Top domain. (b) Subjects.

Figure 2: Confusion matrix showing results for different classification approach.

Overall, the presented approach still faces numerous challenges and leaves room for improvement. One is that there are different versions of DDC available so that the development of a classification system remains challenging, especially for LLMs without specific training. The relatively poor quality of the reference data that merely partially covers the DDC, currently also hinders the improvement of this approach. Although, different bibliographic sources and libraries provide data aligned with DDC classes, they are often skewed to specific domains or subjects, such as in the above dataset, where more than 75 % of the records belong to the Science domain.

5 Conclusion & Future Work

In the presented study, we used large and small language models to classify research metadata based on the DDC. The results show that finetuned SLMs still easily outperform generic LLMs for the classification of research data. We also discussed the main challenges that hinder LLM performance such as different variants of the DDC classification system. However, based on the obtained results, the SLM-based approach has a good potential to be used as an automatic classification model for assigning DDC classes to research data in the near future.

For future work, potential extensions include incorporating more training and test data from more diverse sources, which should also lead to a coverage of additional domains and subjects from the DDC. This could improve the overall performance and ensure a more

balanced distribution of subjects, enhancing generalizability and quality especially with pre-trained SLM-based models.

Data and code availability statement

The replication package, including the code and dataset used in this paper, is openly available at GitHub[5] and is archived[6] (Shahi, Shigapov, and Hummel 2025). If you reuse the dataset, additional attribution to Weber (2019) is needed. The replication package is licensed under the MIT license (for code) and Creative Commons Attribution 4.0 International license (for everything else).

Acknowledgements

This work has been carried out under the TransforMA project, which has received funding from the federal-state initiative "Innovative Hochschule" of the Federal Ministry of Education and Research (BMBF) in Germany.

Authorship Contributions

- Gautam Kishore Shahi: Writing – original draft, experiments;
- Renat Shigapov: Writing – review & editing;
- Oliver Hummel: Writing – review & editing.

Conflict of Interest

Beyond the funding declared under Acknowledgements, all authors have no conflict of interest to declare.

5 https://github.com/TransforMA-WP3/LLM4DDC; *Visited by the editors on June 30, 2025.*

6 https://www.doi.org/10.7801/479 or https://madata.bib.uni-mannheim.de/479; *Visited by the editors on June 30, 2025.*

Bibliography

Church, Kenneth Ward. 2017. “Word2Vec”. *Natural Language Engineering* 23 (1): 155–162. https://doi.org/10.1017/s1351324916000334.

Devlin, Jacob, 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),* 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423.

Dewey, Melvil. 1899. *Decimal Classification and Relativ Index for Libraries: Clippings, Notes, Etc.* Library bureau.

Dewey, Melvil, Joan S. Mitchell, Julianne Beall, Rebecca Green, Giles Martin, and Michael Panzer. 2011. *Dewey decimal classification and relative index.* 23rd edition. Dublin, Ohio: OCLC Online Computer Library Center.

Golub, Koraljka. 2021. “Automated Subject Indexing: An Overview”. *Cataloging and Classification Quarterly* 59 (8): 702–719. https://doi.org/10.1080/01639374.2021.2012311.

Ho, Clara Wan Ching, Tobias Weber, Thorsten Fritze, and Thomas Risse. 2024. “Towards Multilingual LLM-Based Approaches for Automatic Dewey Decimal Classification”. In *International Conference on Theory and Practice of Digital Libraries,* 23–33. Springer, Springer Nature Switzerland. ISBN: 9783031724404. https://doi.org/10.1007/978-3-031-72440-4_3.

Hu, Linmei, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. “Llm vs small model? large language model based text augmentation enhanced personality detection model”. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 38:18234–18242. 16. https://doi.org/10.1609/aaai.v38i16.29782.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* https://doi.org/10.48550/ARXIV.1907.11692.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv. https://doi.org/10.48550/ARXIV.1910.01108.

Shahi, Gautam Kishore, and Oliver Hummel. 2025. “On the Effectiveness of Large Language Models in Automating Categorization of Scientific Texts”. In *Proceedings of the 27th International Conference on Enterprise Information Systems,* 544–554. SCITEPRESS – Science / Technology Publications. https://doi.org/10.5220/0013299100003929.

Shahi, Gautam Kishore, Renat Shigapov, and Oliver Hummel. 2025. *Replication package for "LLM4DDC: Adopting Large Language Models for Research Data Classification Using Dewey Decimal Classification".* Dataset. Visited on June 30, 2025. https://doi.org/10.7801/479. https://madata.bib.uni-mannheim.de/479/.

Wang, Jun. 2009. "An extensive study on automated Dewey Decimal Classification". *Journal of the American Society for Information Science and Technology* 60 (11): 2269–2286. https://doi.org/10.1002/asi.21147.

Watanabe, Yu, Koichiro Ito, and Shigeki Matsubara. 2025. "Capabilities and Challenges of LLMs in Metadata Extraction from Scholarly Papers". In *International Conference on Asian Digital Libraries,* 280–287. Springer. https://doi.org/10.1007/978-981-96-0865-2_23.

Weber, Tobias. 2019. *s-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Field of Study".* Zenodo. https://doi.org/10.5281/ZENODO.3490396.

Weber, Tobias, Dieter Kranzlmüller, Michael Fromm, and Nelson Tavares de Sousa. 2020. "Using supervised learning to classify metadata of research data by field of study". *Quantitative Science Studies* 1 (2): 525–550. https://doi.org/10.1162/qss_a_00049.

Zhou, Shutian, Zizhe Zhou, Chenxi Wang, Yuzhe Liang, Liangyu Wang, Jiahe Zhang, Jinming Zhang, and Chunli Lv. 2024. "A User-Centered Framework for Data Privacy Protection Using Large Language Models and Attention Mechanisms". *Applied Sciences* 14 (15): 6824. https://doi.org/10.3390/app14156824.