

Abstract

Authors: Gautam Kishore Shahi, Renat Shigapov, Oliver Hummel

LLM4DDC: Adopting Large Language Models (LLMs) for Research Data Classification Using Dewey Decimal Classification (DDC)

As the volume of research data continues to grow, accurately classifying this data in institutional, national, and international repositories remains a significant challenge. While the Dewey Decimal Classification (DDC) system is widely used for automatic subject indexing [1] in the context of libraries, its application to automating the creation of metadata for research data is merely a fledgling discipline [2]. This work addresses this gap by evaluating the use of LLMs in automating the detection of research areas for DDC classification of research data. 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.

Our current focus is placed on 3-digit DDC classification, the level of granularity that can maintain interpretability without overwhelming complexity. We evaluate several state-of-the-art models, including Llama 3.1, fine-tuned BERT-like models, and ChatGPT, to determine their effectiveness in performing DDC-based classification. The models were tested on a diverse dataset of research metadata spanning various scientific domains. We also experimented with different prompt-engineering strategies or adjusting parameters such as model temperature for performance optimization. The evaluation was conducted using standard F1-score, precision, and recall. Additionally, we conducted an error analysis to understand the types of misclassifications made by the models and to identify areas for improvement.

In conclusion, our study shows the feasibility and potential of LLMs for automating the classification of metadata of research data using DDC by carefully selecting model parameters and leveraging prompt-engineering strategies from zero-shot and few-shot prompts. The performance of LLMs will be evaluated using precision, recall, and F1-score. To facilitate adoption, we will openly release our models, codes, and data, providing institutions with the necessary tools to integrate LLM-based classifiers into their existing data infrastructures.

References

- [1] Golub, K. (2021). Automated Subject Indexing: An Overview. *Cataloging & Classification Quarterly*, 59(8), 702–719. <https://doi.org/10.1080/01639374.2021.2012311>
- Golub, K. (2021). Automated Subject Indexing: An Overview. *Cataloging & Classification Quarterly*, 59(8), 702–719. <https://doi.org/10.1080/01639374.2021.2012311>

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