


ANNIKA BUCHHOLZ¹ IMENE KHEBOURI² VU THI HUONG³
TIM KUNT⁴ THORSTEN KOCH⁵ WOLFGANG PETERS-KOTTIG⁶
TOMASZ STOMPOR⁷ JANINA ZITTEL⁸

Detecting and classifying publications based on their abstracts with LLM embeddings and multi-label classifiers

1	 0009-0008-3235-0896
2	 0000-0002-2389-8795
3	 0009-0007-4869-0505
4	 0009-0006-5732-3208
5	 0000-0002-1967-0077
6	 0000-0003-4486-2422
7	 0000-0002-7279-2962
8	 0000-0002-0731-0314









Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30 84185-0
Telefax: +49 30 84185-125

E-mail: bibliothek@zib.de
URL: <http://www.zib.de>

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Detecting and classifying publications based on their abstracts with LLM embeddings and multi-label classifiers

Annika Buchholz ¹, Vu Thi Huong ^{1,4}, Imene Khebouri ¹, Thorsten Koch ^{2,3}, Tim Kunt ¹, Wolfgang Peters-Kottig ⁵, Tomasz Stompor ⁵, and Janina Zittel ³

¹Digital Data and Information for Society, Science, and Culture,
Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany

²Software and Algorithms for Discrete Optimization,
Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany

³Applied Optimization, Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany

⁴Institute of Mathematics, Vietnam Academy of Science and Technology, 10072 Hanoi, Vietnam

⁵Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV),
Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany

April 12, 2026

1 Introduction

The ever-growing volume of publications makes organizing, analyzing, and navigating them increasingly difficult. Automated methods promise to ease this workload, usually by providing a form of pre-selection or ranking. The information contained in a citation network built on the citation data of the publications is a prime candidate for such methods:

Citations infer topical proximity, familiarity, and the lineage of science. They induce a simple binary and asymmetric relation (where a *cites* b), which serves as input to algorithms on graphs. Many well-established methods, such as the Leiden algorithm [Traag et al., 2019] and PageRank [Brin and Page, 1998, Chen et al., 2023], operate on this premise.

Texts of the publications themselves would naturally contain the largest share of information; however, they are shrouded by semantics and not as easily parsed into a machine-readable form. This challenge is addressed by natural language processing (NLP), most recently through advances via large language model (LLM) embeddings. Beyond retrieval and unsupervised methods [Keraghel et al., 2024], LLM-embeddings are inherently well-suited for supervised classification. In this work, we built classification methods for publications based on their abstracts from the Web of Science [Clarivate, n.d.], via LLM-embeddings. These classifiers predict the topic of an unknown publication, enabling both the automated labeling of large datasets and the analysis of citation-linkage of semantically proximate publications.

2 Method

LLM-embeddings map texts of any given length and content to a single point in a high-dimensional vector space. This mapping can capture even smaller semantic differences, as it is based on the comprehensive training of the underlying model on large bodies of text. As a result, the output, a simple vector of length usually between 768 and 4096, is small enough

to allow handling datasets from thousands to millions of data points, such as the publication database Web of Science (WoS). If we compare this format to that of a citation network, the resulting set of vectors can be interpreted in multiple ways:

- Instead of a binary, asymmetric relation (such as a cites b), each metric in the vector space gives us a precise number of the distance between two publications a and b . In this way, we obtain a numerical representation of the similarity or closeness of a and b .
- All vectorized abstracts combined reveal a self-structured landscape of texts, a map of sciences, as described in our previous work [Kunt et al., 2025].
- Embedding vectors serve as fitting input for prevalent classification and prediction methods from machine learning, both as a data type and due to them containing the semantic contents in a compressed, yet expressive form.

Citations and proximity in the embedding space, however, encode two fundamentally different relationships. Citations reflect influence, academic genealogy, peers and institutions; in general, the structures and most importantly time and progress. Proximity in the embedding space reflects semantic similarity, which in turn describes shared language, shared topics, and shared views. Proximity in the citation network and proximity in the embedding space correlate (as illustrated by fig.2), but they are not equivalent.

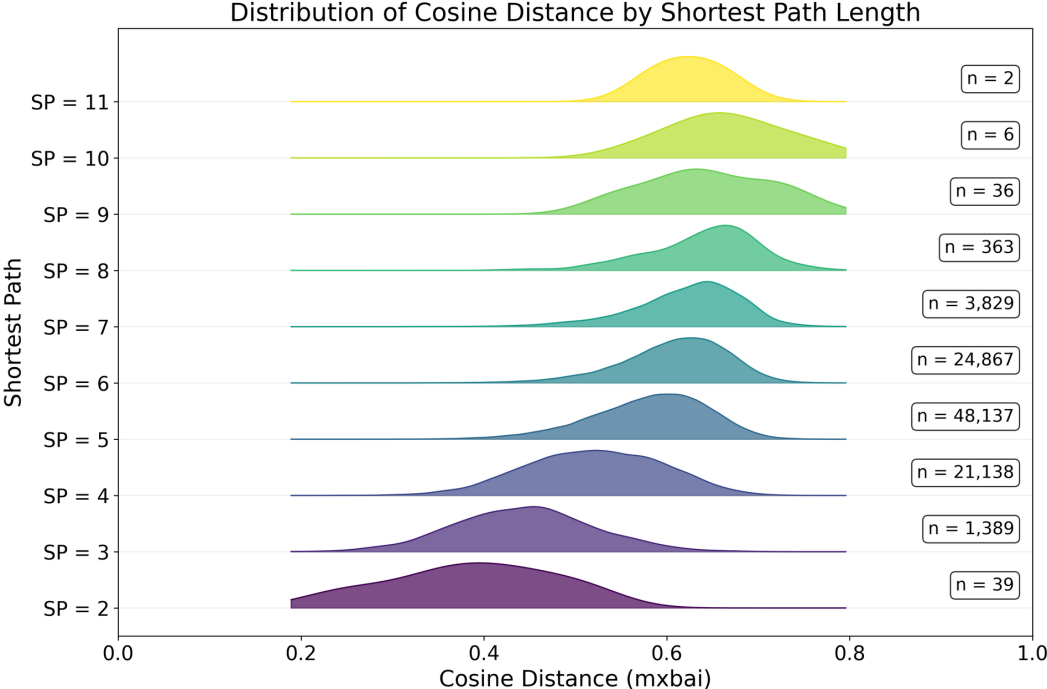


Figure 1: Positive correlation of the distances in the embedding space and on the citation graph (see also [Kunt et al., 2025]): For this plot, we randomly sampled 100,000 pairs of publications from the Web of Science dataset and computed the shortest path on the complete citation graph, as well as the cosine distance in the embedding space.

3 Experiments

Building on our previous study [Kunt et al., 2025], we consider the abstracts of 416,540 publications from WoS and their embeddings. Full-texts are not available for all publications. We also argue that embedding the abstracts serves as a good compromise between computational resources and expressiveness. Abstracts explicitly summarize the intent, scope, methods, and conclusions of their publication. This descriptive quality makes them a well-suited text form for LLM-embeddings.

We use the *mixtral-embed-large:335m* embedding model with 1,024 dimensions [Mixedbread AI, 2024] inside an ollama framework. For a discussion of the model choice and comparison with *nomic-embed-text:v1.5* [Nomic AI, 2024], we refer to our previous study.

As input for training, we choose the *traditional subject* labels by WoS. These labels assign each publication one or more subjects from a list of 255, as classified by WoS. Note that most publications are assigned to multiple subjects, and many of the subjects appear to have overlaps and/or are contained in one another.

With the embeddings and the subject labels set, we can implement supervised learning methods. For this, the different subject labels serve as input for multi-output regression models. We run a ridge regression as a strong baseline and linear model, in addition to a random baseline for reference. Further, we implement two multilayer perceptron (MLP) architectures with Keras to grasp more complex and non-linear relationships between the data points:

- **MLP I** contains an input layer and a single hidden densely-connected ReLU layer of 100 neurons, followed by an output layer with 254 neurons and softmax activation. We use Adam optimizer and evaluate by classification accuracy. Training runs for 100 epochs, however we also run MLP I with early stops to avoid overfitting and reduce unnecessary training time.
- **MLP II** contains four hidden densely-connected ReLU layers with 512 neurons each. To avoid overfitting and accelerate training, we employ $L2$ regularization and batch normalization. MLP II uses Adam too, but adds a learning rate scheduler. We train for max. 200 epochs with early stopping enabled.

MLP I constitutes a simple network architecture, MLP II scales and serves as a test to exhaust the capabilities of this approach or to verify that a simpler model may be adequate. Both the regression model and the MLPs are built to output a vector of subjects, where each potential subject is assigned a weight between 0 and 1. This expresses both the confidence with which the models predict the subject and its belongingness to one or multiple subjects. For all models, the data is divided into a 9 : 1 train-test split. As a comparison and to detect overfitting, we run the same models on a smaller sample of 41,901 publications. Computations run on an Intel Xeon CPU E3-1270 v6 (3.80GHz) with a Nvidia Quadro P400.

4 Discussion

Assessing the predictive qualities of the proposed models merely by numbers, i.e., by evaluating measures of accuracy, hardly does this analysis justice, or as [Held et al., 2021] states: *“Topics as collectively shared perspectives on knowledge cannot be used as ground truths for assessing the outcomes of bibliometric topic reconstruction exercises because the former is an interpretive scheme embedded in human consciousness and the latter is a structured set of publications.”* A thorough qualitative analysis, however, exceeds the scope of this extended

abstract and will be the subject of the accompanying talk.

The following table summarizes the results of our supervised learning approach, comparing the different methods by computational time and error metrics on the test-set. For mean squared error (MSE), we compare the predicted subject vectors on the test set with the WoS *traditional subject* labels. Further, we compute the mean reciprocal rank (MRR), which compares the list of predicted subjects descending by the value assigned by the classifier to the list of *traditional subject* labels. Best performing classifiers for each measure are highlighted in bold.

Sample Size	Classifier	Time	MSE	MRR
41,901	Random Baseline	-	0.0028	0.0278
	Ridge Regression	153.66s	0.0023	0.6249
	MLP I	467.98s	0.0025	0.5867
	MLP I (early stop)	98.20s	0.0020	0.6431
	MLP II (early stop)	165.64s	0.0021	0.6427
416,540	Random Baseline	-	0.0027	0.0297
	Ridge Regression	66.31s	0.0022	0.6361
	MLP I	5517.09s	0.0020	0.6593
	MLP I (early stop)	934.92s	0.0020	0.6672
	MLP II (early stop)	21772.79s	0.0019	0.6766

Due to the clear structure of the data points in the embedding space, the regression baseline already performs as a strong linear predictor. The MLP models outperform across the evaluated metric MRR. Note that results of MLP II on the smaller dataset are included for completeness, however showcases overfitting, as is to be expected at this size.

A qualitative analysis reveals that when the predictors deviate from the WoS labels, they mostly agree on the broader topic but differ in a finer distinction to the field, especially for subjects that overlap. For instance, consider the labels *Multidisciplinary Physics*, *Applied Physics*, and *Physics of Fluids and Plasmas*, all of which are parts of *Physics* but mutually intersect. In such cases, the predicted labels and WoS labels may differ. Likewise, we observe that the performance differs between branches of science, with certain subjects in the humanities proving harder to distinguish from one-another. Many of the mismatching prediction labels appear to be improvements or corrections on the original labeling. This is particularly apparent for publications, mislabeled in the WoS database due to their journal-level labeling scheme categories [Clarivate, 2025]. The classifiers proved to provide better estimates, for the aforementioned publications, of the corresponding subjects over WoS. To conclude, the experiments showcased that the automated classification scheme based on validated training labels is a robust method for structuring publication data to a finer degree. Our future works aim at scaling these experiments on the entire WoS dataset, as well as comparing the results to an implementation on the OpenAlex database [Priem et al., 2022].

Acknowledgements

This work is co-funded by the European Union (European Regional Development Fund EFRE, Fund No. STIIV-001) and supported by the German Competence Network for Bibliometrics (Grant No. 16WIK2101A).

References

- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7), 107–117.
- Chen, Y., Koch, T., Zakiyeva, N., Liu, K., Xu, Z., Chen, C.-h., Nakano, J., & Honda, K. (2023). Article’s scientific prestige: Measuring the impact of individual articles in the web of science. *Journal of Informetrics*, 17(1), 101379.
- Clarivate. (n.d.). *Web of Science*. Retrieved January 2026, from <https://www.webofscience.com>
- Clarivate. (2025). *Web of Science Categories*. Retrieved January 2026, from https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Web-of-Science-Categories?language=en_US
- Held, M., Laudel, G., & Gläser, J. (2021). Challenges to the validity of topic reconstruction. *Scientometrics*, 126(5), 4511–4536.
- Keraghel, I., Morbieu, S., & Nadif, M. (2024). Beyond words: A comparative analysis of llm embeddings for effective clustering. *International Symposium on Intelligent Data Analysis*, 205–216.
- Kunt, T., Buchholz, A., Khebouri, I., Koch, T., Litzel, I., & Vu, T. H. (2025). Embedding large-scale graph and text-based datasets with llms.
- Mixedbread AI. (2024). *Mxbai-embed-large-v1*. Retrieved January 2026, from <https://www.mixedbread.com/docs/models/embedding/mxbai-embed-large-v1>
- Nomic AI. (2024). *Nomic Embed*. Retrieved January 2026, from <https://docs.nomic.ai/api/embeddings-and-retrieval/text-embedding>
- Priem, J., Piwowar, H., & Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.
- Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific reports*, 9(1), 1–12.