Patent Classification with Multi-Vector Embeddings

Giulio Lanza[1,2], Jacopo Staccioli[1,3], Marco Grazzi[3], and Daniele Moschella[1]

[1]Scuola Superiore Sant'Anna, Pisa, Italy
[2]Univ. of Pisa, Italy
[3]Università Cattolica del Sacro Cuore, Milano, Italy

April 2025

References

[1] G. Aslanyan and I. Wetherbee. Patents phrase to phrase semantic matching dataset. *arXiv preprint arXiv:2208.01171*, 2022.

[2] H. Bekamiri, D. S. Hain, and R. Jurowetzki. PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT. *Technological Forecasting and Social Change*, 206:123536, 2024.

[3] F. Cariaggi, C. De Nobili, and S. Bratières. Patents for industrial pollution prevention and control. Technical Report JRC126541, Publications Office of the European Union, 2021.

[4] L. Dhulipala, M. Hadian, R. Jayaram, J. Lee, and V. Mirrokni. MUVERA: Multi-vector retrieval via fixed dimensional encodings. *arXiv preprint arXiv:2405.19504*, 2024.

[5] A. Haghighian Roudsari, J. Afshar, W. Lee, and S. Lee. PatentNet: multi-label classification of patent documents using deep learning based language understanding. *Scientometrics*, 127(1):207–231, 2022.

[6] O. Khattab and M. Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[7] J.-S. Lee and J. Hsiang. Patent classification by fine-tuning BERT language model. *World Patent Information*, 61:101965, 2020.

[8] S. Li, J. Hu, Y. Cui, and J. Hu. DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2):721–744, 2018.

[9] T. J. Lybbert and N. J. Zolas. Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. *Research Policy*, 43(3):530–542, 2014.

[10] M.-T. Nguyen, N. Bui, M. Tran-Tien, L. Le, and H.-T. Vu. CinPatent: Datasets for patent classification. *arXiv preprint arXiv:2212.12192*, 2022.

[11] E. Sharma, C. Li, and L. Wang. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*, 2019.

[12] T. Tran and R. Kavuluru. Supervised approaches to assign Cooperative Patent Classification (CPC) codes to patents. In *International conference on mining intelligence and knowledge exploration*, pages 22–34. Springer, 2017.

[13] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, and Y. Matsumoto. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics, 2020.

[14] I. Yamada and H. Shindo. Neural attentive bag-of-entities model for text classification. *arXiv preprint arXiv:1909.01259*, 2019.

[15] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.