Automatic classification of historical texts using a BERT model: news about wild berries, 1860-1910

Matti La Mela, Uppsala University, matti.lamela@abm.uu.se

Ekta Vats, Uppsala University

This paper presents, applies, and evaluates a method of automatic classification of historical texts based on machine learning and transformer models. The article uses the BERT language model (Bidirectional Encoder Representations from Transformers), which is a powerful and state-of-the-art model for performing various tasks in natural language processing (NLP) such as named-entity recognition and linking, text generation, sentence prediction, or classification (e.g. Labusch & Neudecker 2022; Jiang et al. 2021). The aim of the paper is to examine the performance of the model in classifying historical OCRd texts by applying it in as part of a real research example, while only limited empirical cases using BERT in historical contexts are currently available (e.g. Ardanuy et al. 2020). The benefit of BERT models is that they are multilingual, or built with national language data (Haffenden et al. 2023; Virtanen et al. 2019; Aprosio et al. 2022), which allows to employ the classification method in other linguistic or even multilingual contexts.

The empirical case presented in the paper concerns the classification of historical newspaper articles, which are a common source used in digital history research. The paper examines a corpus of texts about wild berries from Finnish 19th century newspapers, and classifies them according to how the news media reported about the use of the wild berries as an economic resource. The long-term changes in the article types are used to understand the process of commodification of wild nature since the late nineteenth century.

The paper proceeds through the three steps of the classification method: 1) building the training data, 2) classification of newspaper articles with BERT language model, and 3) evaluation of the results through error analysis and the case study.

First, training data was created manually for the model by a domain expert. In total, 415 newspaper articles were annotated by dividing them into descriptive-narrative articles and economic-industrial articles. The former includes, for instance, descriptions of the nature or news pieces where berry-pickers appeared as actors (e.g. snake bite a berry picking child). The latter includes articles that regard economic and industrial use of wild berries in local, national, or international contexts (e.g. exports of wild berries or advertisement).

Second, the trained model is used to classify the complete corpus of newspaper articles about berry-picking. The current study corpus consists of 1805 newspaper articles about wild berry picking published in the Finnish newspaper in 1860-1910. The newspaper articles have been extracted from the DIGI-collection of the National Library of Finland (https://digi.kansalliskirjasto.fi/). The OCR quality of the newspaper articles in the DIGI collection are on 70-75 % (Kettunen and Pääkkönen 2016). For the classification task, the

paper employs Simple Transformers NLP library and Finnish-language BERT models (Virtanen et al. 2019). The Python code and documentation is available at https://github.com/ektavats/BerryBERT.

Third, we evaluate the results of the classification with an evaluation sample, and conduct an error analysis of the results by close reading them. The manual evaluation is currently on-going, but already by using a small sample of the training data (10%) we reach a very high F1-score of 0.97. This is probably due to the binary classification and the homogeneous nature of the corpus. At the same time, as illustrated in Figure 1, the errors in the classification enable us to identify that the binary classification should be expanded into more categories, which are currently border cases or "hidden" as part of "descriptive-narrative" class.

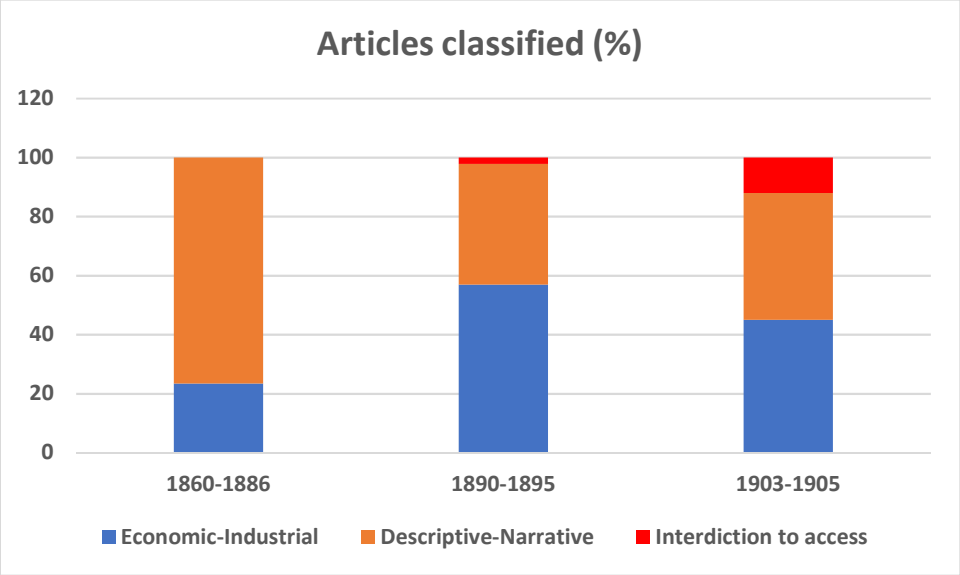

Figure 1. Emerging third category "Interdictions" as part of the descriptive-narrative articles.

Finally, to illustrate the usability of the classification, we use place name information that has been previously recognized in the articles part of the wild berry corpus to visualize the classifications (For the named-entity recognition process, see La Mela et al. 2019). As shown in Figure 2, we see that the descriptive-narrative about wild berry picking are about smaller localities mainly situated in Finland. The locations in the economic-industrial articles are much more concentrated and regard larger centers and also country names. These geographic differences highlight the content of the articles. For example, export destinations and port cities are visible in the economic-industrial articles.

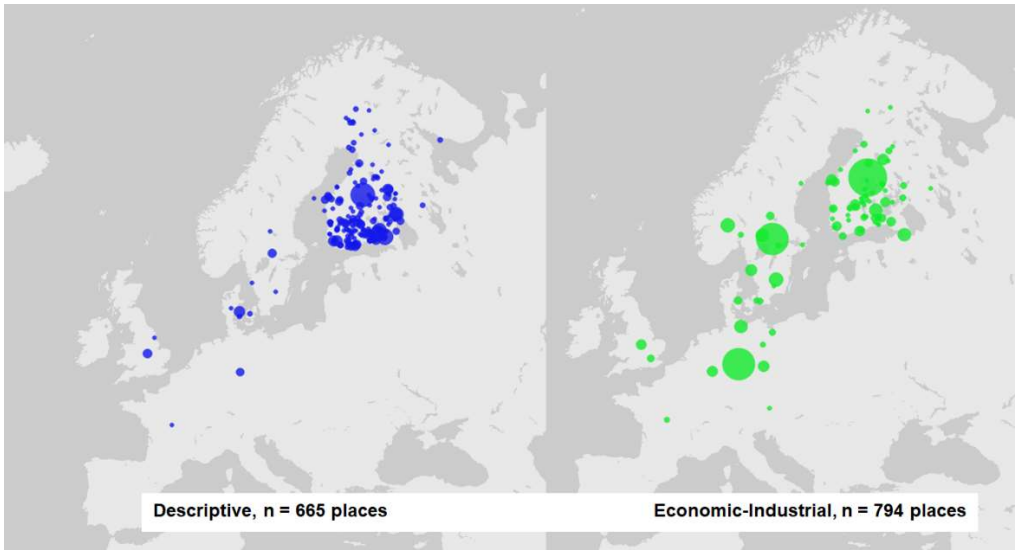

Figure 2. Place names identified in the descriptive-narrative articles (left), and the economic-industrial articles (right), 1903-1905.

The paper concludes that we are able to apply in a purposeful way a BERT language model that is built with contemporary training data for classifying historical OCRd newspaper texts. This is probably related to our use of binary classes and rather homogeneous and distinctive text material. Jiang et al. (2021) have shown that BERT performs well also when studying OCRd and sometimes messy historical texts especially when the model is fine-tuned for the context and the data analysed. Moreover, the paper suggests that the use of machine learning requires attention on the evaluation of the tool itself, but also that the errors and challenging prediction cases can lead to important insights and modifications of the research setting. In this case, the classification helped us to discover an emerging third category as part of the descriptive-narrative articles.

References

Aprosio, A. P., Menini, S., and Tonelli, S. 2022. BERToldo, the Historical BERT for Italian. In Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pages 68–72, Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lt4hala-1.10.pdf

Ardanuy, M. C., Nanni, F. Beelen, K. et al. 2020. Living Machines: A study of atypical animacy. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4534–4545, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://aclanthology.org/2020.coling-main.400

Haffenden, C., Fano, E., Malmsten M., and Börjeson, L. 2023 [2021]. Making and Using AI in the Library: Creating a BERT Model at the National Library of Sweden. College & Research Libraries, accepted for publication (January 2023).

Jiang, M., Hu, Y., Worthey, G., Dubnicek, R. C., Underwood, T. J., and Downie, S. 2021. Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book

Excerpts. Proceedings of the Conference on Computational Humanities Research 2021 Amsterdam, the Netherlands, November 17-19, 2021. CEUR-WS vol. 2989, 266-279. https://ceur-ws.org/Vol-2989/long_paper43.pdf

Kettunen, K. and Pääkkönen, T. 2016. Measuring Lexical Quality of a Historical Finnish Newspaper Collection — Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 956–961, Portorož, Slovenia. https://aclanthology.org/L16-1152.pdf

La Mela, M., Tamper, M., and Kettunen, K. 2019. Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8. CEUR Workshop Proceedings, vol. 2364, 295-307. http://ceur-ws.org/Vol-2364/

Labusch, K., and Neudecker, C. 2022. Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT. Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum. CEUR-WS, https://ceur-ws.org/Vol-3180/paper-85.pdf
Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. 2019. Multilingual is not enough: BERT for Finnish. arXiv:1912.07076 [cs.CL], https://doi.org/10.48550/arxiv.1912.07076.