A Decade of DerStandard Data

large-scale longitudinal data, signed networks, NLP, semantic embeddings, user comments

Extended Abstract

Social media have become a defining aspect of the first quarter of the 21st century. Their active use, now over many years and by many users provides us with the opportunity to collect large-scale data with a strong longitudinal angle. In this work, we release a data set for the DerStandard newspaper platform that covers an entire decade. DerStandard is an Austrian print newspaper that was founded in 1988. As an internet pioneer, already in 1995 it went online and soon after the first chat rooms became active on the page. In further course, these chat rooms developed into a forum where registered users can post below newspaper articles. The forums are highly active with $247,864$ users posting $75,644,850$ comments between the beginning of 2013 and the end of 2022.

For a detailed overview of all records in the data set see Figure 1. The data contains not only large amounts of text in the form of user comments but also other behavioral measures such as up- and down-votes on these user postings ($412,511,165$ votes in total). We include information on which user exactly voted in which way (positive or negative) on which postings. This feature allows the reconstruction of signed networks of online user interaction. In addition, we provide publicly available information on the articles themselves that serve as the basis for the user postings below. This article information includes self-annotated content tags by DerStandard staff. To facilitate the use of textual data and to preserve user privacy, we provide pre-computed vector representations of the more than 75 million texts in the data set. We create these textual embeddings using a state-of-the-art specialized open model ”KaLM-embedding-multilingual-mini-v1”[1] model from the Hugging Face Model Hub [2]. We share all of the data in publicly accessible repositories.

DerStandard is a source of high-quality text as the platform uses semi-automatic moderation schemes [1] to remove automated accounts and unwanted contents for many years already. With the release of the data, we enable researchers to go beyond typical (US)-English centric contributions: We provide a large-scale textual data source in German, a medium-resourced language spoken by around 100 million people worldwide. We expect the data to be of great interest to social scientists, as the period covered by our data features many contentious (political) events happening in the country of Austria. Austria, despite being one of the typical smaller and less influential European nations, is often seen as a model for later, significant developments in Germany and other powerful Western European countries. Its location in Central Europe and its historical role as a crossing point between Eastern, Western and Southern Europe make it share many features of the bordering countries around and an interesting case to study for many social scientific research questions.

The data we provide was used in part previously in three separate studies. The first work featured almost in real-time monitoring of the affective expressions of the entire DerStandard community during COVID-19 [2]. Second, we showed that the results of sentiment analysis conducted on user comments on DerStandard closely track the dynamics of mood explicitly

[1] `https://huggingface.co/HIT-TMG/KaLM-embedding-multilingual-mini-v1`

[2] Commit ID of the exact model version used: 8a82a0cd2b322b91723e252486f7cce6fd8ac9d3

expressed by users in a survey that was run on the platform in the same time period [3]. Third, we extracted signed networks of user interactions on the platform to identify the main divisions between users. This analysis allowed us to pinpoint issues that reinforce societal fault lines and contribute to polarization, as well as topics—such as COVID-19—that spark online conflict without strictly following these dividing lines [4].

We provide technical validations for several key aspects of our dataset. Analyses of vector similarities confirm the quality of the provided embeddings by showing expected patterns in comment proximity. Specifically, similarity scores are highest for comment pairs where one is a reply to the other, followed by lower similarity for comments in the same thread (not necessarily direct responses to each other), and finally, lowest similarity for any pair of comments under the same article. Additionally, we examine similarity distributions for comments under articles with different topic labels versus the same topic, finding that cross-topic pairs generally have lower similarity scores. Among same-topic pairs, more specific labels—such as *football* or *Middle East*—exhibit higher similarities than broader categories like *policy issues*.

We expect this dataset to enable future research in a number of domains. As one example, it can serve as a so-far unique resource in network science. It provides large-scale, longitudinal and real-world ”found” data that directly allows building networks of signed interactions between thousands of users. Previous work in the domain of signed networks either had to work with small-scale data or use proxy measures to assume the positivity or negativity of interactions, for example through Wikipedia edits [5]. On the article level, information such as content tags annotated by DerStandard staff themselves can directly be used as valuable gold standard labels to validate clustering approaches. Furthermore, to make other researchers additionally flexible in the kinds of research questions and methodologies they want to address, we provide the possibility for data enrichment by enabling linkage to the ”One Million Post Corpus” [6] of DerStandard earlier released by the Austrian Research Institute for Artificial Intelligence (OFAI)[3]. The One Million Post Corpus contains full text and manually annotated labels for a subset of the data.

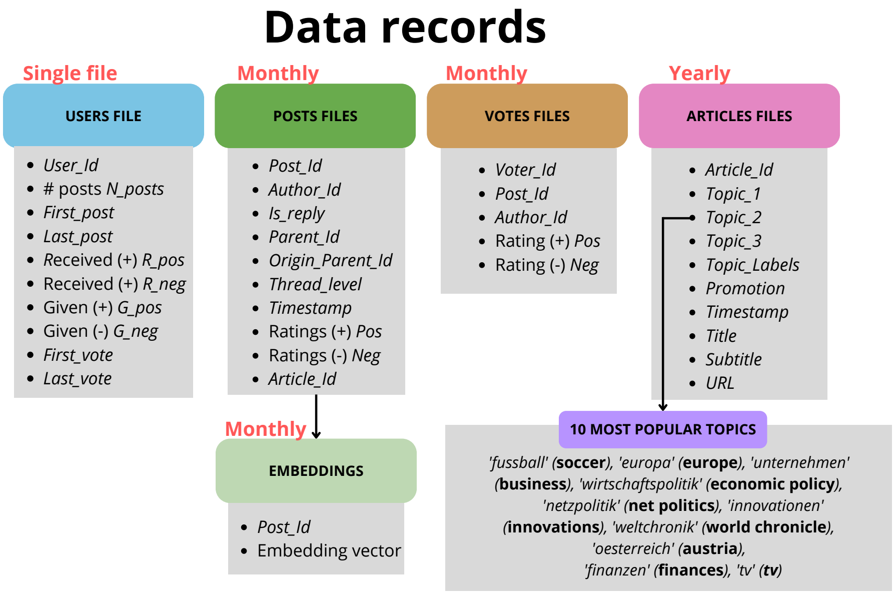

Figure 1: **Summary of all variables provided with the DerStandard data set and the organization in files.**

[3] https://ofai.github.io/million-post-corpus/

References

[1] D. Schabus and M. Skowron, "Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, eds.), (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[2] M. Pellert, J. Lasser, H. Metzler, and D. Garcia, "Dashboard of Sentiment in Austrian Social Media During COVID-19," *Frontiers in Big Data*, vol. 3, 2020.

[3] M. Pellert, H. Metzler, M. Matzenberger, and D. Garcia, "Validating daily social media macroscopes of emotions," *Scientific Reports*, vol. 12, p. 11236, Dec. 2022.

[4] E. Fraxanet, M. Pellert, S. Schweighofer, V. Gómez, and D. Garcia, "Unpacking polarization: Antagonism and alignment in signed networks of online interaction," *PNAS Nexus*, vol. 3, p. pgae276, 07 2024.

[5] S. Maniu, B. Cautis, and T. Abdessalem, "Building a signed network from interactions in Wikipedia," in *Databases and Social Networks on - DBSocial '11*, (Athens, Greece), pp. 19–24, ACM Press, 2011.

[6] D. Schabus, M. Skowron, and M. Trapp, "One Million Posts: A Data Set of German Online Discussions," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Shinjuku Tokyo Japan), pp. 1241–1244, ACM, Aug. 2017.