SentiANNO: Annotating Sentiment in Austrian Historical Newspapers

Krusic, Lucija
lucija.krusic@uni-graz.at
Department for Digital Humanities, University of Graz, Austria

Hochreiter, Clara
cla.hochreiter@gmail.com
Department for Digital Humanities, University of Graz, Austria

Frauendorfer, Melanie
melanie.frauendorfer@uni-graz.at
Department for Digital Humanities, University of Graz, Austria

This contribution presents the preliminary version of SentiANNO, a sentiment-annotated corpus derived from sentences in historical Austrian newspapers, and details the annotation process and the training of annotators. The corpus is intended for fine-tuning existing Machine Learning models for the NLP (Natural Language Processing) task of Sentiment Analysis (SA), facilitating the automatic detection of emotions and opinions in texts (Liu, 2012). This corpus enables SA of journalistic texts in Austrian German from 1800 to 1938 and can be applied to historical texts across various topics.

Currently, there is a significant lack of sentiment-annotated corpora needed for fine-tuning existing Machine Learning models, such as Schweter (2020), which was trained on non-annotated historical newspapers. Despite significant efforts to create sentiment and emotion-annotated corpora of German historical dramas (Schmidt et al., 2021), no comparable resource exists for Austrian historical newspapers.

The base corpus comprises sentences and texts from newspapers available through the ANNO (Österreichische Nationalbibliothek, 2021) and DIGITARIUM (Austrian Academy of Sciences, 2017) collections. It includes texts from newspapers such as Wienerisches Diarium, Das Vaterland, Neue Freie Presse, Arbeiter Zeitung, and Illustrierte Kronen Zeitung, spanning the years 1800 to 1938. This time frame was selected for two reasons: it encompasses significant migration and societal developments during the long 19th century and the pre-, during-, and post-war periods of the early 20th century; and documents published after 1938 are excluded due to copyright restrictions, as these works may still be protected under intellectual property laws.

The corpus was compiled using a two-step methodology: Dynamic Topic Modeling with BERTopic, followed by classification. Annotated with topics such as migration, national minorities, education, and labor, it provides a comprehensive resource for analyzing societal changes during this period.

The base corpus is then annotated with four sentiment categories: positive (positive sentiment is expressed in the sentence), negative (the sentence expresses a negative sentiment), neutral (there is no sentiment in the sentence), and mixed (two sentiments are expressed, it is not possible to find a clear dominant one). The annotations were provided by a team of three semi-expert annotators (Master's students of Linguistics and Digital Humanities). Sentences were used as the unit of annotation, with an average sentence length of 35.7 tokens, the shortest sentence having four tokens and the longest having 350 tokens. A sentence was used as the annotation unit because sentiment often changes within an article and sometimes even within a sentence.

The annotators were trained in sentiment analysis and annotation using an iterative approach, which incorporated established annotation methodologies (Sprugnoli et al., 2023; Schmidt et al., 2018). The annotation process was conducted in stages, with each stage comprising 50 to 150 annotation units centered on texts related to specific topics. Following each stage, annotators provided feedback regarding challenges and ambiguities encountered during the task. This feedback informed subsequent adjustments to the process, leading to the inclusion of additional contextual information to facilitate decision-making. The added context consisted of the newspaper title, the publication date, and the sentences immediately preceding and following the target sentence.

Traditionally, tools such as spreadsheets and word processors have been commonly used for annotation collection (Sprugnoli and Redaelli, 2024; Sprugnoli et al., 2023; Schmidt et al., 2018). However, the selection of annotation tools can significantly impact the quality of annotations. For example, Schmidt et al. (2019) compared various tools, including Word, WebAnno, CATMA, eMargin, and Sentimentator, and found that using a specialized tool like Sentimentator can increase annotator confidence and certainty in their decisions.

A comparative evaluation was conducted to determine the most suitable method for annotation collection (Krušić, 2024a). The tools assessed were Google Forms (a survey platform), Google Sheets (a spreadsheet application), and Doccano (a dedicated annotation tool) (Hiroki et al., 2018). The comparison focused on criteria such as the ease of presenting sentences to annotators, the clarity of sentence and annotation visualization, navigability between sentences, and the annotators' ability to provide comments. Doccano was identified as the most effective and easy-to-use annotation tool (Krušić, 2024a). This comparison will be extended in future annotation rounds, thus including other annotation tools, e.g. INCEpTION (Klie et al, 2024).

Annotator feedback highlighted several challenges, including the complexity of the historical language and context, metaphors and satire, and the expression of multiple sentiments in one sentence (Krušić, 2024a). Despite these complexities, preliminary results indicate a fair to moderate level of agreement (0.405), which is in line with expectations and previous work (Sprugnoli et al., 2023; Schmidt et al., 2019, 2018). Further, the annotation process yielded an Average Percentage Agreement (APA) of 92.5% for majority agreement (achieved by at least two out of three annotators). These outcomes demonstrate the reliability of the annotation methodology, supporting the development of a gold-standard corpus consisting of 930 sentences, with plans for further expansion in subsequent annotation phases.

The SentiAnno corpus, characterised by high linguistic and historical variation, promises significant contribution for the fine-tuning of models in the context of Digital Humanities (DH). Furthermore, the comprehensive annotation guidelines, along with the insights and experiences accumulated throughout the annotation process, will serve as valuable resources for the DH community. The preliminary corpus is available in the SentiAnno Github repository (Krušić, 2024b). Further annotation rounds will be conducted in the scope of the project "SentiAnno: Sentiment-annotated corpus of Austrian historical newspapers", funded by CLARIAH-AT. The extended corpus will be made publicly available on Zenodo, a platform that supports open access and aligns with the FAIR principles.

Bibliographie

Austrian Academy of Sciences. (2017, 2023). Wien[n]erisches Diarium. DIGITARIUM. https://digitarium-app.acdh.oeaw.ac.at/start.html?id=jg17xx

Klie, J.-C., Eckart de Castilho, R. and Gurevych, I. 2024. "Integrating INCEpTION into larger annotation processes". Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) - System demonstrations.

Krušić, Lucija. 2024a. "Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience". In Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 51–62, Miami, USA. Association for Computational Linguistics.

Krušić, Lucija. 2024b. "SentiAnno [GitHub repository]". GitHub.https://github.com/lucijakrusic/SentiAnno

Liu, B. 2012. "Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)". Morgan & Claypool Publishers.

Österreichischen Nationalbibliothek. 2021. "ANNO (AustriaN Newspaper Online)". anno.onb.ac.at

Schmidt, T., Burghardt, M., and Dennerlein, K. 2018." Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior". In Proceedings of the Workshop on Annotation in Digital Humanities co-located with ESSLLI 2018 (annDH 2018) (pp. 47-52). Sofia, Bulgaria. DOI: urn:nbn:de:bvb:355-epub-437018.

Schmidt, T., Dennerlein, K., and Wolff, C. 2021. "Towards a Corpus of Historical German Plays with Emotion Annotations". 11 pages, 741719 bytes. https://doi.org/10.4230/OASICS.LDK.2021.9

Schweter, S. 2020. "Europeana BERT and ELECTRA models (1.0.0)". https://doi.org/10.5281/zenodo.4275044

Sprugnoli, R., Mambrini, F., Passarotti, M., and Moretti, G. 2023. "The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace". Italian Journal of Computational Linguistics, 9(1). https://doi.org/10.4000/ijcol.1125