Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience

Lucija Krušić
Department of Digital Humanities
University of Graz
lucija.krusic@uni-graz.at

Abstract

This study presents the development of a sentiment-annotated corpus of historical newspaper texts in Austrian German, addressing a gap in annotated corpora for Natural Language Processing in the field of Digital Humanities. Three annotators categorised 1005 sentences from two 19th-century periodicals into four sentiment categories: positive, negative, neutral, and mixed. The annotators, Masters and PhD students in Linguistics and Digital Humanities, are considered semi-experts and have received substantial training during this annotation study. Three tools were used and compared in the annotation process: Google Sheets, Google Forms and Doccano, and resulted in a gold standard corpus. The analysis revealed a fair to moderate inter-rater agreement (Fleiss' kappa = 0.405) and an average percentage agreement of 45.7% for full consensus and 92.5% for majority vote. As majority vote is needed for the creation of a gold standard corpus, these results are considered sufficient, and the annotations reliable. The study also introduced comprehensive guidelines for sentiment annotation, which were essential to overcome the challenges posed by historical language and context. The annotators' experience was assessed through a combination of standardised usability tests (NASA-TLX and UEQ-S) and a detailed custom-made user experience questionnaire, which provided qualitative insights into the difficulties and usability of the tools used. The questionnaire is an additional resource that can be used to assess usability and user experience assessments in future annotation studies. The findings demonstrate the effectiveness of semi-expert annotators and dedicated tools in producing reliable annotations and provide valuable resources, including the annotated corpus, guidelines, and a user experience questionnaire, for future sentiment analysis and annotation of Austrian historical texts. The sentiment-annotated corpus will be used as the gold standard for fine-tuning and evaluating machine learning models for sentiment analysis of Austrian historical newspapers with the topic of migration and minorities in a subsequent study.

1 Introduction

Sentiment analysis (SA), the automatic identification of attitudes, opinions, and emotions in textual data, has been popular since the early 2000s (Liu, 2012). Deriving from Natural Language Processing (NLP), it was initially used to study contemporary data, including reviews and microblog posts. Recently, it has gained prominence in Digital Humanities (DH), expanding beyond contemporary texts to historical and literary texts (Häußler and Gius, 2023; Koncar et al., 2020; Kim and Klinger, 2019). However, texts such as newspaper articles, novels, letters, and poetry, which are commonly studied in DH, pose a challenge due to their formal structures and historical nuances, making sentiment analysis difficult (Kaur and R. Saini, 2014).

Traditional dictionary-based SA methods, heavily relied upon in DH, involve annotating words and phrases with sentiment values. Although this method is easily interpretable and transparent, sentiment dictionaries suffer from low reusability and do not consider word context, missing nuances such as sarcasm or negation (Schmidt et al., 2021c; Schmidt and Burghardt, 2018). To address these limitations, context-sensitive transformer-based machine learning models such as BERT have been developed (Devlin et al., 2018) (Suissa et al., 2022). These models require less annotated data data than traditional ML algorithms (such as BOW or TF-IDF), since they can be pre-trained on large, unannotated datasets. This means that pre-training is usually done once and the model can then be further fine-tuned for various specific purposes and tasks, such as e.g. named-entity recognition or sentiment analysis, using a smaller annotated corpus (of e.g. sentences, plays or verses with corresponding sentiment annotation). This is particularly use-

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 51–62
November 16, 2024.

ful in DH, where the annotation process is often tedious and time-consuming due to the complexity of literary and historical texts.

However, good quality annotations are crucial for the accuracy of the models with which they are fine-tuned. To better understand the conditions necessary for creating a high-quality sentiment-annotated fine-tuning corpus for texts in DH, recent studies have focused on the annotation process (Al-Laith et al., 2024; Sprugnoli et al., 2023), annotator behavior (Schmidt et al., 2018), and annotation tools (Schmidt et al., 2019). These studies discuss the relevance of expert versus non-expert annotators, the optimal tool for sentiment annotation, and the importance of guidelines. These insights informed the approach taken in this study, as will be shown in the next chapters.

There is still a gap in sentiment-annotated DH corpora that could be used for fine-tuning Machine Learning models, one such model being presented in Schweter (2020). This model was trained on non-annotated historical newspapers and offers the possibility of further fine-tuning with an annotated corpus, for a specific task - such as named-entity recognition or sentiment analysis. However, newspaper texts pose various annotation challenges: historical language and context, discriminatory language, sarcasm and metaphors.

The following sections present the current state of sentiment annotation in Digital Humanities and describe the creation of a sentiment-annotated corpus of Austrian historical newspapers through an annotation study. The annotations, the annotation process and annotation tools are evaluated quantitatively and qualitatively. This evaluation identifies key challenges and provides guidelines for annotating historical newspaper texts. These initial results will guide future refinements of the corpus, which will be openly accessible on Zenodo in accordance with the FAIR principles

2 Sentiment annotation for the Digital Humanities

Sentiment analysis (SA), in the context of Digital Humanities (DH), has often been used to answer specific research questions related to literary or historical studies. It has been used to analyse German plays (Schmidt et al., 2021a), Spanish Enlightenment periodicals (Koncar et al., 2021), Spanish song lyrics (Hernández-Lorenzo et al., 2022) and conflict in German novels (Häußler and Gius, 2023). Texts investigated by DH, such as newspapers, novels, poetry, and drama, present unique challenges due to their formal structures and historical linguistic nuances, making SA and annotation particularly complex (Kaur and R. Saini, 2014).

Sprugnoli et al. (2023) lists several aspects of the annotation process that are to be considered when constructing an annotation study: classification granularity, type of annotator (expert, non-expert or crowd workers), perspective, unit of annotation, and language of annotation unit. A further consideration (Schmidt et al., 2019) is the choice of annotation collection tool, which can have an effect on the annotation experience. These factors can impact the inter-rater agreement, which informs about the quality of annotations and is crucial in the development of a gold standard corpus.

In SA, two primary classification tasks are typically addressed: polarity and emotion analysis. Polarity analysis focuses on determining the direction of the sentiment within the text, often classified into categories such as positive, negative, and neutral (Liu, 2012). For more complex analyses, polarity classification may involve finer distinctions, such as differentiating between highly negative and highly positive sentiments, often using a numerical scale or additional categories. Sprugnoli et al. (2023) annotate four categories: positive, negative, neutral, and mixed. On the other hand, emotion analysis refers to the classification into emotional categories, often following Ekman's theory of basic emotions (Ekman, 1992) or Russell's circumplex model (Russell, 1980). Schmidt et al. (2019) conduct a polarity annotation study of German historical plays by G.E. Lessing, using extended polarity categories - negative, positive, neutral, mixed, uncertain, and other. In a subsequent study (Schmidt et al., 2021b), they conducted an additional annotation study with emotion categories. In these studies, inter-rater agreement, measured by statistical measures of Cohen's or Fleiss' kappa and Krippendorff's alpha, decreases with the higher number of categories that need to be annotated (Sprugnoli and Redaelli, 2024). However, these studies show that the agreement on literary and historical texts ranges from poor to moderate agreement, due to subjectivity and difficulty of the annotation process.

When conducting sentiment annotation, experts are preferred annotators, due to their accuracy and deep understanding of complex texts (Sprugnoli et al., 2023). But, they are scarce and expensive

(Schmidt et al., 2018). Semi-experts, such as advanced students, provide a more accessible alternative with reasonably reliable results (Yeruva et al., 2020; Schmidt and Burghardt, 2018), while non-experts (e.g. obtained through crowd-sourcing), though less accurate, can be effectively utilised in large-scale projects with appropriate guidance and annotation schemes (Schmidt et al., 2018)).

Sentiment annotation can focus on two perspectives: the emotions the author intended to convey, or the emotions perceived by the reader (Sprugnoli and Redaelli, 2024). Most studies focus on sentiments as intended by the author of the text (Sprugnoli et al., 2023; Häußler and Gius, 2023; Schmidt et al., 2019), as the annotation from the perspective of the reader can lead to low inter-rater agreement, due to subjectivity of the task.

The unit of annotation is also significant, and highly dependent on the type of text. Sprugnoli et al. (2023) and Häußler and Gius (2023) annotate sentences, while Schmidt et al. (2018) annotate speeches in a larger play. Annotating a shorter unit can be beneficial, as it minimises the change in sentiment shifting within the annotation unit.

Traditionally, spreadsheets and Word have often been used as the main tool for annotation collection (Sprugnoli and Redaelli, 2024; Sprugnoli et al., 2023; Schmidt et al., 2018). Schmidt et al. (2019) compare various annotation tools, such as Word, WebAnno, CATMA, eMargin and Sentimentator. They report using Sentimentator (a dedicated annotation tool) and Word increases annotator levels of certainty, thus, making the choice of annotation tool important for obtaining high-quality annotations. They employ standard usability and user experience questionnaires, NASA-TLX (Hart and Staveland, 1988) and User Experience Questionnaire (UEQ-S) (Hinderks et al., 2018), to assess user experience and perceived annotator workload.

With respect to the previous work presented in this section, the annotators in this study are semi-experts who have received extensive training in sentiment annotation. They annotate sentences in four categories: positive, negative, neutral and mixed, focusing on the sentiment intended by the writer. Furthermore, Google Forms (an online survey tool), Google Sheets (an online spreadsheet tool) and Doccano (Hiroki et al., 2018) are compared for the annotation process in order to establish an optimal tool for future annotation processes. To assess the usability of different annotation tools, Google Forms, Google Sheets, and Doccano (Hiroki et al., 2018) were compared. The findings from this comparison, along with insights into annotator experiences, will inform the choice of tools for future sentiment annotation projects. Additionally, user experience was assessed using a combination of NASA-TLX, UEQ-S, and custom questions targeting the specific challenges of annotating historical texts.

3 Aims and research questions

The main aim of this study is establishing the optimal conditions for sentiment annotation of Austrian historical newspapers, with the goal of creating a reliable gold standard corpus for fine-tuning of ML models for sentiment analysis. This study aims to answer the following questions:
RQ1 Is using semi-expert annotators appropriate for the task of annotation of historical newspapers?
RQ2 How does the historical language and context of the texts influence the annotation process?
RQ3 How do the annotators perceive the difficulty of the annotation task?
RQ4 Which tool is most optimal for sentiment annotation of historical newspapers?

4 Methods

4.1 Corpus

The corpus used for the annotation consisted of 1005 sentences from two Austrian periodicals, “Neue Freie Presse” and “Das Vaterland”. The newspapers were published between 1850 and 1900. The corpus was created using Dynamic Topic Modelling with BERTopic (Grootendorst, 2022)), and through this process was automatically annotated with topics such as "migration", "labour", "Jews", "Croats", “Czechs”, etc. 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.

4.2 Annotation process

The corpus was annotated by three semi-expert annotators (Masters and PhD students in Linguistics and Digital Humanities), two native German speakers and one fluent German speaker. The annotators were previously familiar with the task of sentiment

analysis and received additional training for the task of sentiment annotation. Each annotator was individually introduced to the corpus and the annotation process, followed by practical examples. The annotators were assigned to the annotation tasks for 3 months, 5 hours per week.

They individually annotated the sentiment in four categories:

- Positive (positive sentiment is expressed in the sentence)
- Negative (the sentence expresses a negative sentiment)
- Neutral (there is no sentiment in the sentence)
- Mixed (two sentiments are expressed, it is not possible to find a clear dominant one).

With regard to annotation perspective (Sprugnoli et al., 2023) they annotated the sentiment the author intended to convey through the sentence. The process was organised in stages, with group meetings after each round of annotation to exchange feedback, provide further training and resolve any uncertainties. In the first stage, 50 comments were provided in the form of a Google Forms survey with multiple choice questions. No additional information was provided in this round. In the second stage, 232 annotation units were provided to the annotators via a spreadsheet in Google Sheets. This time, in response to annotator feedback, the previous and subsequent sentences were provided as additional context, as well as the name of the journal and the date of publication. A column for comments was also added so that annotators could leave comments about their annotation choices if they felt it was necessary.

In the following 5 annotation rounds, they were given the remaining 723 sentences, divided into separate annotation tasks. The sentences were annotated using Doccano, an open source data labelling tool for machine learning tasks such as classification (Hiroki et al., 2018). Doccano was built in Python using the Django library, and an instance of it was deployed using Heroku for this annotation study. Doccano allows the upload and download of datasets in various non-proprietary formats (including csv). The main benefits are the ease of assigning annotation units to users, the ability to view one annotation unit at a time and navigate between them, and the ability to view additional information about the annotation unit on the side of the screen. You can also easily track your progress, adding a gamification aspect to the annotation process. This has previously been shown to be beneficial to the user experience (Schmidt et al., 2019).

4.3 Evaluation

The annotations were evaluated for inter-rater agreement using Fleiss' Kappa (McHugh, 2012) and average percentage agreement (APA) overall and per category.

At the end of all annotation tasks, annotators were asked to complete a questionnaire (administered via Google Forms) about various aspects of the annotation process and the annotation tools used. The questionnaire comprised seven sections and a total of 25 questions. The complete questionnaire is available in Appendix A. The questionnaire included questions on the overall perceived difficulty of the annotation task, the perceived time taken to complete the annotations, and confidence in the annotations. The impact of historical language, context and specific linguistic features (such as sarcasm and metaphor) on the complexity of the annotation process was also examined. A section dedicated to the comparison of annotation tasks evaluated the ease of use of each tool and the speed of adaptation to the tool using a 5-point Likert scale.

In addition, following the recommendation of (Schmidt et al., 2019), two standardised usability tests were used to quantify the overall usability and user experience, namely the Nasa Task Load Index (NASA-TLX) (Hart and Staveland, 1988) and the User Experience Questionnaire (UEQ-S) (Hinderks et al., 2018). The NASA-TLX assesses the perceived workload of a task across different dimensions such as mental, physical and temporal (Schmidt et al., 2019). The scores are then combined and averaged into an overall workload score. The UEQ-S (User Experience Questionnaire - Short Version) (Hinderks et al., 2018) is used to quantitatively assess user experience across two key dimensions: Pragmatic Quality (PQ) and Hedonic Quality (HQ). PQ evaluates the usability and functionality of a product or task, indicating how easy it is for users to accomplish their goals. HQ measures the emotional appeal, reflecting how enjoyable, engaging, and motivating the product or task is for users. By analysing both PQ and HQ scores, the UEQ-S provides a comprehensive

overview of the user experience. Using the above metrics and questionnaire, the quality of annotation and the experience of annotators will be assessed.

5 Annotation evaluation

The inter-rater agreement, as measured by Fleiss' kappa, indicates a fair to moderate level of agreement (0.405), which is in line with expectations. The Average Percentage Agreement (APA) is 45.7% when all three annotators concur. However, for a sentence to be included in the gold standard, a majority vote – where two out of three annotators agree – is deemed sufficient. This resulted in a 92.5% agreement among at least two annotators across all annotation items in our corpus, which was used to create a gold standard of 930 items. Of the total number of annotations, 447 were classified as negative, 345 as neutral, 81 as positive, and 56 as mixed. The results indicate that the token length may have an impact on the annotations. Items rated as 'mixed' had a mean of 53 tokens, compared to means of 31, 29, and 38 tokens for 'positive', 'neutral', and 'negative' annotations, respectively. This indicates that additional splitting of the annotation items may be required to eliminate ambiguity. Notably, two annotators classified approximately half of the units as negative, while the third annotator rated 34% as negative and 45% as neutral, indicating that the most predominant sentiments were negative and neutral (see Table 1).[1].

By closely examining the sentences in the annotation study, we can identify the causes of disagreement. For instance, the sentence "In den Beziehungen zwischen Polen und Czechen ist, trotz der vielen gegenseitigen Versicherungen brüderlicher Freundschaft, in jüngster Zeit — wie fast regelmäßig vor jedem Wiederzusammentritte des Reichsrathes — eine Spannung eingetreten." (English translation: "In the relations between Poles and Czechs, despite the many mutual assurances of brotherly friendship, a tension has recently arisen—almost always before each reconvening of the Reichsrat.") is highly ambiguous. It conveys both positive and negative sentiments between the subjects (Czechs and Poles) and suggests a possible disdain by the author ("almost always"). In contrast, sentences with complete agreement, such as "Treibt auch Noth einen Serben oder Walachen zur Arbeit, so strengt er sich durchaus nicht an." (English translation: "Even if necessity drives a Serb or Wallachian to work, they do not exert themselves at all."), present a clearer sentiment and a more evident object of that sentiment.

6 Evaluation of annotator experience and annotation tools

The questionnaire yielded valuable insights into the attitudes and perceptions of the annotators regarding the annotation process. In terms of the complexity of the text annotation process, two participants rated it as "challenging" (4 on a 5-point scale), while one rated it as "moderate" (3 on a 5-point scale). The estimated time required for one annotation unit is between one and four minutes, with one annotation round taking between three and over five hours. Two annotators reported needing to take regular breaks from the annotation process. However, two out of three annotators report having high confidence (4 out of 5 on a Likert scale) in their annotations, while one reports moderate confidence (3 out of 5). In terms of specific difficulties, annotating longer texts was identified as particularly challenging, particularly when sentences were complex and required close attention. Another challenge identified was the need to remain objective, as one participant mentioned the difficulty of not letting personal beliefs influence the annotation process. The historical context and language of the texts had a notable impact on the annotation process. The participants indicated that the historical context affected their ability to annotate, with one annotator finding it particularly challenging. This emphasizes the importance of familiarity with the historical background when conducting sentiment analysis.

Moreover, all participants felt that the historical context significantly influenced their ability to annotate sentiment, reporting the need to independently research the historical background using resources like Wikipedia, the University Library Catalogue, Britannica, and the ANNO repository. The historical language, including vocabulary, grammar, and phrasing, also posed challenges, similarly influencing their ability to annotate sentiment.

Participants indicated that the clarity of what should be annotated—whether it was the sentiment of the language, sentiment towards a group, or the emotional state of the speaker—was not always clear. They reported that regular discussions and feedback sessions were useful for overcoming these

[1] A Jupyter Notebook outlining the evaluation can be found at https://github.com/lucijakrusic/SentiAnno/

Annotation	Annotator 1	Annotator 2	Annotator 3
negative	550 (54.7%)	435 (43.33%)	346 (34.36%)
neutral	236 (23.51%)	423 (42.13%)	453 (45.12%)
positive	147 (14.61%)	73 (7.27%)	79 (7.87%)
mixed	71 (7.07%)	73 (7.27%)	126 (12.55%)

Table 1: Annotation results across three annotators.

challenges.

The use of sarcasm and metaphors in the texts presented a challenge for most participants. Two respondents found metaphors challenging, while one was affected by sarcasm, indicating a need for additional training or guidelines on handling figurative language in sentiment analysis.

Access to the previous and following sentences (context) was generally seen as helpful, with all participants agreeing that it aided in making more accurate annotations. However, opinions were divided on whether more context was necessary, with one participant suggesting that additional context could clarify ambiguous sentiments.

The overall NASA-TLX score was calculated by averaging the scores on six dimensions: mental, physical, temporal demand, performance, effort and frustration level. This equates to a score of 3.09 out of 5, indicating that the perceived workload is slightly above neutral. The annotators indicated that the level of workload was moderate overall, with some dimensions (such as mental demand, effort, and performance) rated higher than others. In terms of the UEQ-S, the Pragmatic Quality (PQ) achieved a score of 4.97 on a scale from 1 to 7 (1 reflects a poor experience in terms of usability, and a higher score of 7 reflects a good experience). This indicates that the task was moderately usable and clear. The hedonic quality (HQ) was rated at 4.45 out of 7 (with 1 indicating a less enjoyable experience and 7 indicating an emotionally satisfying one), indicating that the task was perceived as somewhat enjoyable and not unpleasant.

The questionnaire also assessed the usability of different annotation tools, with participants evaluating Google Forms, Google Sheets, and Doccano. All three annotators identified Doccano as the most intuitive tool, citing its clear layout, ease of navigation, and effective display of context as key factors. Additionally, the ability to leave comments and track progress was identified as a valuable feature. However, both Google Forms and Google Sheets are also considered relatively straightforward to use and easily adaptable (see Figures 1 and 2). The annotators found Google Sheets less practical for navigating between the annotation units and viewing the full sentences. This is reflected in Figure 1, where one annotator noted that Google Sheets were difficult to use.

Lastly, annotators provided constructive feedback on how to improve the annotation process. One suggestion was to standardise the token length of context provided for each sentence, as inconsistencies sometimes made interpretation difficult. Another recommendation was to allow annotators to "correct" incomplete sentences by adding parts from adjacent sentences. These insights will be implemented in the following rounds of annotation.

7 Discussion

In this study, a sentiment-annotated corpus of Austrian historical newspaper texts was developed, with three semi-expert annotators categorizing sentences into four sentiment classes: positive, negative, neutral, and mixed. The inter-rater agreement, measured using Fleiss' kappa, resulted in a score of 0.405, indicating fair to moderate agreement and reflecting the inherent challenges in annotating historical corpora. This aligns with previous research ((Sprugnoli et al., 2023; Schmidt et al., 2019, 2018)) and highlights the difficulty of classifying complex texts, particularly those with mixed sentiments. This category, indicating both positive and negative sentiments within a single sentence, was the most challenging due to its higher token count, and presumably as a result, higher content ambiguity. This finding underscores the complexity of historical texts, where sentiments can shift within the same sentence or be expressed through nuanced language, including sarcasm and metaphors, that is difficult to categorise definitively. The study also observed a notable imbalance in sentiment categories, with the majority of annotations marked as negative or neutral. This distribution mirrors the historical context of the periodicals, which frequently adopted a critical stance toward migration

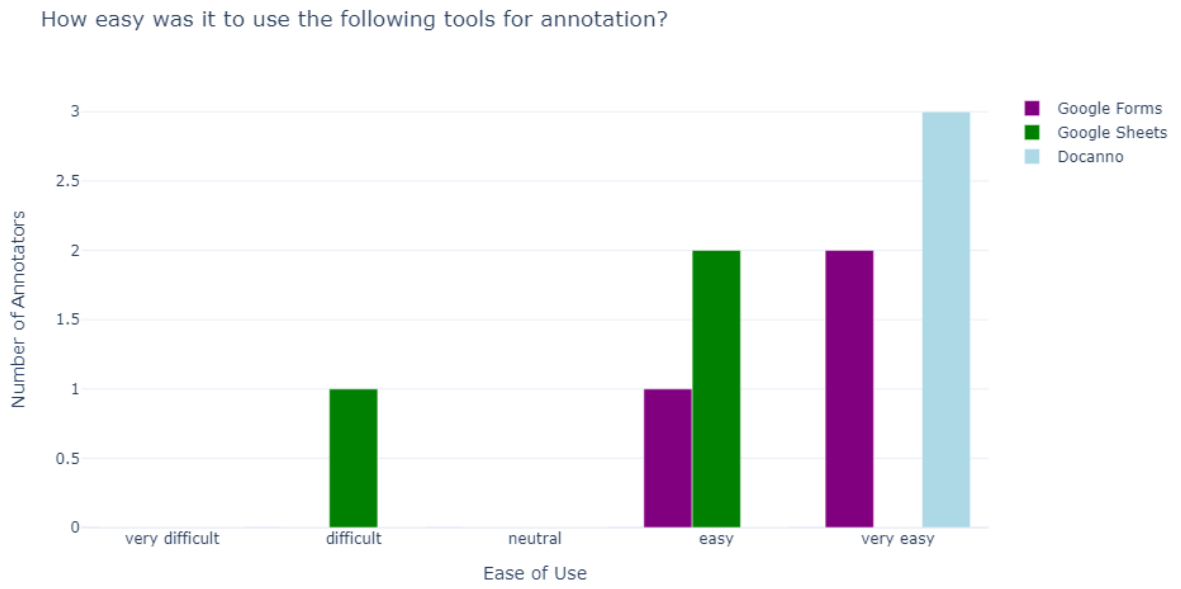

Figure 1: Annotation tool comparison on ease of use.

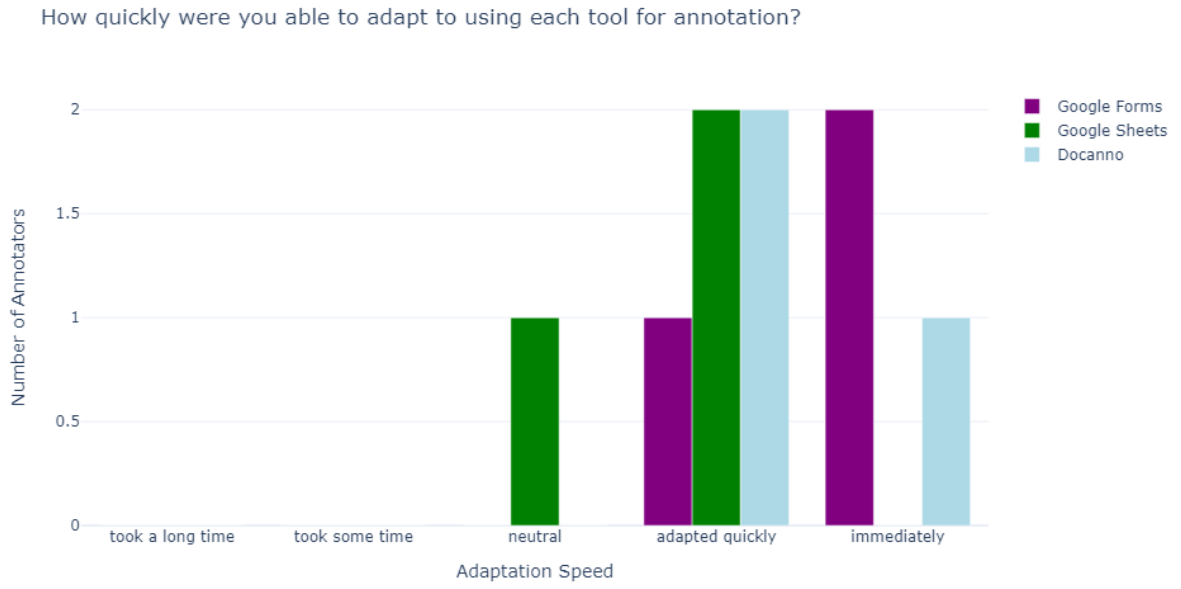

Figure 2: Annotation tool comparison on speed of adaptation.

and minorities. Despite these challenges, the annotation process achieved an Average Percentage Agreement (APA) of 45.7% for full agreement and 92.5% for majority vote (two out of three annotators). These results validate the reliability of the annotation process, allowing the creation of a gold standard corpus comprising 930 sentences, which will be extended in future annotation rounds.

Notably, the successful use of semi-expert annotators—advanced students—demonstrates that it is possible to achieve reliable annotations without relying on fully trained experts. This finding corroborates previous studies (Yeruva et al., 2020; Schmidt and Burghardt, 2018), reinforcing the notion that semi-experts, can serve as an accessible yet effective alternative for similar tasks.

The annotation process was systematically supported by providing essential contextual informa-

tion, including references to the previous and following sentences, newspaper name, and date, which helped annotators interpret complex historical texts more accurately. Such contextual cues were particularly important given the nuanced language found in historical materials, where sentiments often shift within a sentence. The user experience questionnaire (Appendix A) revealed the cognitive and emotional demands on annotators, especially when dealing with historical language, figurative expressions like metaphors, and sarcasm. This feedback is valuable for improving future annotation workflows and provides a basis for comparing annotator experiences in similar tasks. Reusing the questionnaire, particularly its sections on historical language and context, could further enhance the systematic evaluation of annotation processes within Digital Humanities projects.

The standardized usability assessments, NASA-TLX and UEQ-S, highlighted the need to consider both cognitive workload and user engagement when designing annotation tasks. The NASA-TLX results showed that while the task was manageable, it required significant cognitive effort, particularly for complex, sentiment-laden historical texts. This finding aligns with Schmidt et al. (2019) and highlights the importance of considering workload when designing annotation tasks, particularly for complex historical texts. The UEQ-S results reveal a clear process (Pragmatic Quality) but suggest the task could be more engaging (Hedonic Quality). While Doccano proved to be the most user-friendly tool, with a positive impact on annotator efficiency and accuracy, there is room for improvement in user experience, particularly regarding task engagement. These results reinforce the need for comprehensive guidelines and tool evaluations, as well as attention to annotator workload, to ensure efficient and accurate sentiment annotation in Digital Humanities.

8 Conclusion

This study contributes to the field of Digital Humanities by presenting the first sentiment-annotated corpus of Austrian historical newspaper texts in Austrian German. Through the collaboration of three semi-expert annotators, 930 sentences were annotated for sentiment using a carefully designed process supported by tools like Doccano, Google Sheets, and Google Forms. The fair-to-moderate inter-rater agreement (Fleiss' kappa of 0.405) reflects the challenges of annotating historical texts, where sentiment is often complex and contextually dependent.

A contribution of this study is the user experience questionnaire, which were specifically designed to assess the cognitive and emotional challenges encountered during the annotation process. The bespoke sections of the questionnaire not only provided valuable insights for improving subsequent annotation rounds but also offer a reusable framework for evaluating annotator experiences in other historical and literary annotation projects.

Furthermore, this study highlights the feasibility of employing semi-expert annotators in sentiment annotation, achieving reliable results through thorough guidelines and iterative feedback. Standardized assessments of usability and user experience, combined with the custom questionnaire, provided critical insights into annotators' cognitive demands and areas where the task could be improved.

By making the corpus openly available, this research offers a valuable resource for further sentiment analysis in Austrian German, particularly on topics such as migration, minorities, and labor rights. The findings and methodology outlined here will serve as a basis for future annotation projects, contributing to more nuanced and accessible sentiment analysis in historical and literary contexts.

Limitations

It should be noted that this study is subject to a number of limitations. Firstly, the limited number of annotators may impact the representativeness of the findings, particularly in terms of inter-rater agreement.

Secondly, the imbalance in sentiment categories, with a predominance of negative and neutral annotations, may have had an impact on the overall results. This imbalance reflects the content of the newspapers, but it also presents a challenge for model training and evaluation, as models may be biased towards these more common categories. Further rounds of annotation will be added to the corpus in the future, with the aim of reducing this imbalance.

Thirdly, the historical context and language of the texts presented significant challenges to the annotators, who had to navigate complex sentences and cultural references that may not have been immediately apparent. While the annotators were semi-experts, additional training or the use of annotators with expertise in history or media studies

could help to overcome some of the challenges identified.

Furthermore, the annotators indicated that they would have benefited from additional context accompanying the annotation unit and the ability to correct over-split annotation units, which will be addressed at a future stage in the annotation process.

The process of annotation of further data from historical newspapers will continue (with an extension of the temporal coverage and the addition of other newspapers with different political leanings). These limitations can serve as lessons that can be applied in the future to improve the creation of the gold standard.

Ethics Statement

This study was conducted with careful consideration of ethical principles, particularly in relation to the sensitive nature of historical newspaper content. The annotation units in this study included topics such as migration, labour, and minorities, which are often associated with discriminatory language and sentiments. The annotators were instructed to approach these texts with sensitivity and objectivity, ensuring that their annotations reflect the sentiment expressed in the text rather than their personal beliefs or biases.

Moreover, the historical context of the texts was acknowledged as a potential source of bias, both in the content of the texts themselves and in the interpretation by the annotators. To mitigate this, the annotators were provided with extensive training and were encouraged to research the historical background of the texts using reputable sources.

The study also adhered to ethical guidelines regarding the use of human participants. The annotators were informed about the purpose of the study, the tasks they were required to perform, and the potential challenges they might face. Their participation was voluntary and paid.

The study recognizes the potential impact of creating a sentiment-annotated corpus for historical texts, particularly in terms of how these texts may be interpreted and used in future research. The authors are committed to ensuring that the corpus is used responsibly and that any findings derived from it are presented in a manner that respects the historical context and the individuals represented in the texts.

Acknowledgements

We would like to express our sincere gratitude to Melanie Frauendorfer, Clara Hochreiter and Leona Münzer for their diligence and hard work on the corpus annotation. Their insights and feedback are greatly appreciated.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv:2203.05794 [cs].

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. 52:139–183.

Laura Hernández-Lorenzo, Aitor Diaz, Alvaro Perez, Salvador Ros, and Elena González-Blanco. 2022. Exploring Spanish contemporary song lyrics through Digital Humanities methods: Some thematic and structural properties. *Digital Scholarship in the Humanities*, 37(3):738–746.

Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2018. A Benchmark for the Short Version of the User Experience Questionnaire:. In *Proceedings of the 14th International Conference on Web Information Systems and Technologies*, pages 373–377, Seville, Spain. SCITEPRESS - Science and Technology Publications.

Nakayama Hiroki, Kubo Takahiro, Kamura Junya, Taniguchi Yasufumi, and Liang Xu. 2018. {doccano}: Text Annotation Tool for Human.

Julian Häußler and Evelyn Gius. 2023. Operationalizing and Measuring Conflict in German Novels. In *Proceedings of the Computational Humanities Research Conference (CHR 2023)*, volume 3558.

Jasleen Kaur and Jatinderkumar R. Saini. 2014. Emotion Detection and Sentiment Analysis in Text Corpus: A Differential Study with Informal and Formal Writing Styles. *International Journal of Computer Applications*, 101(9):1–9.

Evgeny Kim and Roman Klinger. 2019. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *Zeitschrift für digitale Geisteswissenschaften*. ArXiv: 1808.03137.

Philipp Koncar, Alexandra Fuchs, Elisabeth Hobisch, Bernhard C. Geiger, Martina Scholger, and Denis Helic. 2020. Text sentiment in the Age of Enlightenment: an analysis of spectator periodicals. *Applied Network Science*, 5(1):33.

Philipp Koncar, Bernhard C. Geiger, Christina Glatz, Elisabeth Hobisch, Sanja Sarić, Martina Scholger, Yvonne Völkl, and Denis Helic. 2021. A Sentiment Analysis Tool Chain for 18th Century Periodicals: Experimente in den Digital Humanities.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers, Vermont, Australia.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Thomas Schmidt and Manuel Burghardt. 2018. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.

Thomas Schmidt, Manuel Burghardt, and Katrin Dennerlein. 2018. Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior. pages 47–52. Sofia, Bulgaria. Conference Name: Annotation in Digital Humanities (annDH) Meeting Name: Annotation in Digital Humanities (annDH).

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021a. Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 67–79, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021b. Towards a Corpus of Historical German Plays with Emotion Annotations. pages 11 pages, 741719 bytes.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021c. Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. *Fabrikation von Erkenntnis: Experimente in den Digital Humanities* -.

Thomas Schmidt, Brigitte Winterl, Milena Maul, Alina Schark, Andrea Vlad, and Christian Wolff. 2019. Inter-Rater Agreement and Usability: A Comparative Evaluation of Annotation Tools for Sentiment Annotation.

Stefan Schweter. 2020. Europeana BERT and ELECTRA models.

Rachele Sprugnoli, Francesco Mambrini, Marco Passarotti, and Giovanni Moretti. 2023. The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace. *Italian Journal of Computational Linguistics*, 9(1).

Rachele Sprugnoli and Arianna Redaelli. 2024. How to Annotate Emotions in Historical Italian Novels: A Case Study on I Promessi Sposi. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 105–115, Torino, Italia. ELRA and ICCL.

Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. 2022. Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, 73(2):268–287.

Vijaya Kumari Yeruva, Mayanka Chandrashekar, Yugyung Lee, Jeff Rydberg-Cox, Virginia Blanton, and Nathan A Oyler. 2020. Interpretation of Sentiment Analysis with Human-in-the-Loop. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3099–3108, Atlanta, GA, USA. IEEE.

A Appendix A - Annotation questionnaire

A.1 Section 1 - General experience

1. How would you describe the difficulty of annotation of texts? (5 point Likert scale, very easy - very difficult)

2. How much time did you approximately spend on an annotation unit (sentence)? (short answer)

3. How much time did you approximately spend on 150 annotation units - one annotation round? (short answer)

4. Did you need to take frequent breaks from annotation due to the difficulty of the task? (multiple choice - yes/no/other)

5. How would you describe your confidence in your annotations? (5 point Likert scale, not confident at all - really confident)

6. Provided for you were the texts from two Austrian newspapers - "Neue Freie Presse" (NFP)

and "Das Vaterland" (VTL). The name (or the abbreviation) was also provided. Was one of them more difficult to annotate, and if so, which one? (multiple choice - NFP/VTL/not sure/both were equal in difficulty)

7. What did you find most difficult about the annotation process? (short answer)

A.2 Section 2 - Nasa Task Load Index (Hart and Staveland, 1988)

Please rate your experience on the following aspects of the task:

1. How mentally demanding was the task? (5 point Likert scale, very low - very high)
2. How physically demanding was the task? (5 point Likert scale, very low - very high)
3. How hurried or rushed was the pace of the task? (5 point Likert scale, very low - very high)
4. How successful were you in accomplishing the task? (5 point Likert scale, very low - very high)
5. How hard did you have to work to accomplish your level of performance? (5 point Likert scale, very low - very high)
6. How insecure, discouraged, irritated, stressed and annoyed were you? (5 point Likert scale, very low - very high)

A.3 Section 3 - User Experience Questionnaire (UEQ-S) (Hinderks et al., 2018)

Please rate your experience on the following aspects of the task:

1. Annoying - enjoyable (7 point Likert scale, left extreme - right extreme)
2. Not understandable - understandable (7 point Likert scale, left extreme - right extreme)
3. Slow - fast (7 point Likert scale, left extreme - right extreme)
4. Unpleasant - pleasant (7 point Likert scale, left extreme - right extreme)
5. Complicated - easy (7 point Likert scale, left extreme - right extreme)
6. Boring - exciting (7 point Likert scale, left extreme - right extreme)
7. Demotivating - motivating (7 point Likert scale, left extreme - right extreme)
8. Difficult to learn - easy to learn (7 point Likert scale, left extreme - right extreme)

A.4 Section 4 - Historical Language and Context

1. How much did the historical context of the texts affect your ability to annotate the sentiment? (5 point Likert scale, not at all - significantly)
2. How much did the historical language (e.g., vocabulary, grammar, phrasing) of the texts affect your ability to annotate the sentiment? (5 point Likert scale, not at all - significantly)
3. Did you feel the need to investigate the historical background of the texts on your own? (multiple choice - yes/no)
4. If yes, which resources did you use for this research (please specify)? (short answer)
5. How clear was it what should be annotated: the sentiment of the language, the sentiment towards a person/group of people, the sentiment towards a subject or the emotional state of the speaker? (5 point Likert scale, completely unclear - very clear)
6. If the task was unclear, what would have/has helped you overcome it? (short answer)

A.5 Section 5 - Specific Language Properties

1. How much did the appearance of sarcasm in the texts affect your annotation? (5 point Likert scale, not at all - significantly)
2. How much did the appearance of metaphors in the texts affect your annotation? (5 point Likert scale, not at all - significantly)
3. How much did having access to the previous and following sentences (context) help you in making accurate annotations? (5 point Likert scale, not at all - significantly)
4. Do you believe more context is necessary? (multiple choice - yes/no/other)

A.6 Section 6 - Tool usability comparison

1. How easy was it to use the following tools for annotation?
   (a) Google Forms (5 point Likert scale, very easy - very difficult)
   (b) Google Sheets (5 point Likert scale, very easy - very difficult)
   (c) Doccano (5 point Likert scale, very easy - very difficult)
   (d) How quickly were you able to adapt to using each tool for annotation?
   (e) Google Forms (5 point Likert scale, took a long time - immediately)
   (f) Google Sheets (5 point Likert scale, took a long time - immediately)
   (g) Doccano (5 point Likert scale, took a long time - immediately)
   (h) Which tool did you find the most intuitive to use for annotation tasks? (multiple choice - Google Forms/ Google Sheets/Doccano)
2. Please shortly elaborate why (short answer)

A.7 Section 7 - Additional feedback

1. This section is for any additional observations and remarks. How would you improve the annotation setup? Do you have any additional feedback or advice on how to improve the annotation process? (long answer)