Modeling Moravian Memoirs: Ternary Sentiment Analysis in a Low Resource Setting

Patrick D. Brookshire
Digital Academy
Academy of Sciences and Literature | Mainz
patrick.brookshire@adwmainz.de

Nils Reiter
Department of Digital Humanities
University of Cologne
nils.reiter@uni-koeln.de

Abstract

The Moravians are a Christian group that has emerged from a 15th century movement. In this paper, we investigate how memoirs written by the devotees of this group can be analyzed with methods from computational linguistics, in particular sentiment analysis. To this end, we experiment with two different fine-tuning strategies and find that the best performance for ternary sentiment analysis (81 % accuracy) is achieved by fine-tuning a German BERT model, outperforming in particular models trained on much larger German sentiment datasets. We further investigate the model(s) using SHAP scores and find that the best performing model struggles with multiple negations and mixed statements. Finally, we show two application scenarios motivated by research questions from religious studies.

1 Introduction

While not entirely uncontroversial (cf. Mortimer, 2002, 189ff.), ego-documents (i.e. documents in which humans write about themselves and their experiences) are an important source of historical research (Burke, 2013; Farbstein, 1998; Kuromiya, 1985; Redlich, 1975). In this paper, we focus on one specific kind of ego-document, often called memoir: semi-autobiographical records written by members of the Moravian Church in the 18th century. In line with general migration movements at that time, many Moravians migrated from Europe to America. The semi-autobiographical records we investigate are the result of a custom among Moravians to document their lives in written form. As they were completed, compiled and collected by other members of the respective local church (Van Gent, 2017), we consider them semi-autobiographical. Religiously, the Moravians are connected to the so-called "Blood and Wounds" theology (Atwood, 2006), dating back to their founder, Nikolaus Ludwig von Zinzendorf (1700–1760). Next to blood, the "wounds Jesus suffered on the cross became the main focus of this religious attention" (Atwood, 2006, 38).

The memoirs are also believed to express a high degree of emotionality (Van Gent, 2017; Faull and McGuire, 2022), which is why we focus on sentiment analysis in this paper, while also taking into account that emotionality found in the text is not necessarily only rooted in emotions of the person the memoir is about. From a linguistic standpoint, these sources exhibit regional variation and domain-specific terms, some of them connected to the "Blood and Wounds" theology. We therefore experiment with multiple ways of assigning sentiment scores, and explore the gain by adapting these systems to the specific domain and text genre. We also investigate how to visualize what these models actually have learned, and provide two application scenarios motivated by research interests from religious and historical studies.

In the following sections, we outline connected fields of research first. Then, we go into details about how we compiled our dataset and what kind of analyses we conducted before discussing our findings.

2 Related Work

This paper has links to multiple research areas from Computational Linguistics (CL) and Digital Humanities (DH).

2.1 Digital Biographical Research

Biographical documents have been investigated in both disciplines for quite some time, often working with Wikipedia data (Biadsy et al., 2008; Palmero Aprosio and Tonelli, 2015; Chisholm et al., 2017) or focusing on digitization and editorial work which is often combined with Linked Open Data (Fokkens et al., 2014; Hyvonen et al., 2019). Target-wise, most works see biographies as factual texts from which facts can be extracted. Thus, there is a prevalence of spatio-temporal and social network

Proceedings of LaTeCH-CLfL 2024, pages 91–100

analysis approaches (Faull, 2021; Windhager et al., 2017). To the best of our knowledge, there are only a few other studies that focus on emotions or sentiment in this area. One concerned Australian World War I diaries (Dennis-Henderson et al., 2020) and another one English Moravian memoirs from the 18th century (Faull and McGuire, 2022; McGuire, 2021). The latter is our main reference project.

2.2 Historical Sentiment Analysis

Sentiment analysis is a common CL task that has mainly been applied to news, product reviews and Social Media data (cf. Liu, 2012). Nevertheless, the number of studies devoted to historical domains increased over the past decade. An earlier application was a dictionary-based analysis of relationships between characters in Shakespeare's plays (Nalisnick and Baird, 2013). The former prevalence of sentiment dictionaries gave also rise to corpus-based domain adaptation methods that use seed lists (Hamilton et al., 2016). Regarding historical German data, one of the earliest approaches was a happy ending prediction based on a support vector machine (Zehe et al., 2016). More recent studies found that transformers outperform other approaches (Schmidt et al., 2021; Allaith et al., 2023). However, custom dictionaries are still used in particular for highly specific research questions or small heterogeneous datasets which are typical features of ego-document collections (Faull and McGuire, 2022; Dennis-Henderson et al., 2020).

2.3 Explainable AI

While dictionary-based approaches to sentiment analysis are inherently explainable, transformer-based ones are not, which raises questions about their trustworthyness. This is why various ways to explain a given model globally or locally (i.e. in relation to an individual prediction) have been proposed (Danilevsky et al., 2020; Linardatos et al., 2020). One such method relies on an architecture that not only predicts class labels but also summarizes its input (Bacco et al., 2021). A different one is SHAP (SHapley Additive exPlanations) which is an external explainability model that unifies several similar methods (Lundberg and Lee, 2017). Zielinski et al. (2023) evaluated it on a (non-historical) sentiment analysis application where it performed best with regard to BERT models in terms of plausibility and faithfulness.

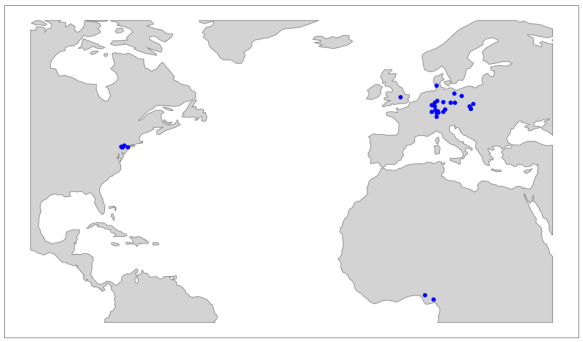

Figure 1: Birthplaces documented in the entire dataset

3 Data and Methodology

In this section, we first describe how we selected the data we wanted to analyze. Afterwards, we list the sentiment and explainability models we used and outline how we conducted our experiments.

3.1 Corpus Construction and Annotation

Due to the lack of German Moravian corpora with sentiment annotations, we needed to compile a new one ourselves. We started with 41 Moravian memoirs transcribed by Faull[1] and added 23 from the crowdsourcing project Moravian Lives[2]. This project lists 328 more documents as available but unpublished since their respective transcriptions are incomplete. Even more Moravian texts from various genres are in the process of digitization but not considered here (Lasch, 2023). To the best of our knowledge, there are at the moment only 64 German memoirs available in digital form, all of which include metadata about the person's gender as well as birth and death dates. This metadata was manually enriched by the place of birth and used to semi-randomly select a gender balanced subcorpus of 36 texts from people that lived in the 18th century. Figure 1 shows that people from various German speaking communities in Europe and New York/Pennsylvania are included but also a few Native Americans and two former African slaves.

Having selected our subcorpus, we first conduct sentence splitting using stanza (Qi et al., 2020). Since punctuation is often not normalized (or not used at all) in the data, we corrected the sentences semi-automatically, which yields 2210 sentences in total. Afterwards, we anonymized each sentence by masking names with `{NAME}` and annotated it with one of the three labels `negative`, `neutral` and

[1] https://katiefaull.com/moravian-materials
[2] http://moravianlives.org/

Dataset	Instances			
	neg	neut	pos	**Total**
Train	485	523	760	1768
Test	115	150	177	442
Total	600	673	937	2210

Table 1: Train/test dataset statistics

positive, making it a ternary classification task. It should be noted that the neutral class was used for sentences without sentiment bearing words, and not for mixed-sentiment sentences. In cases of mixed sentiment, we based our annotations on the final state or result of the action described in a given instance. For example, we annotated (1) as negative.

(1) Ich versuchte oft und viel mir selbst aus diesem Zustand zu helfen, aber vergebens,
('I tried often and hard to help myself out of this state, but in vain,')

Finally, we randomly split our data in 80 % training samples and 20 % used for testing. Both datasets show a positive bias (see Table 1) unlike the prevalence of negative samples in some literary corpora (Allaith et al., 2023; Schmidt et al., 2021).

3.2 Ternary Sentiment Analysis

In our experiment, we compare i) trained off-the-shelf models for sentiment analysis, ii) dictionary-based methods and iii) models fine-tuned to this specific dataset. We use the following systems:

ger-senti-bert. This BERT model was trained on 1.8M German samples from Social Media as well as app, hotel and movie reviews (Guhr et al., 2020).

senti-distilbert. This Hugging Face model[3] was distilled from the zero-shot classification pipeline on the Multilingual Sentiment dataset[4].

SentiWS. This dictionary lists polarity values within the interval [-1, 1] for 34.6k German word forms. It has a focus on financial data and product reviews (Remus et al., 2010).

GerVADER. VADER (Valence Aware Dictionary for sEntiment Reasoning) adds a few context-aware rules to a dictionary lookup (Hutto and Gilbert, 2014). The German adaptation builds on word forms from SentiWS as well as a few slang words all of which are re-rated in a crowdsourcing project and enriched with items commonly found in Social Media (Tymann et al., 2019).

We fine-tuned bert-base-cased (Devlin et al., 2019) and gbert-base (Chan et al., 2020) with the transformers library from Hugging Face using default parameter settings (i.e. 3 epochs). This took approximately 10 minutes on a T4 GPU. We also fine-tuned ger-senti-bert in the same way to evaluate whether this kind of transfer learning is a viable option.

3.3 Experimental Setup

The experiment we conduct is a **sentence-wise classification** experiment, to determine which of the systems/models mentioned above performs best. To this end, we transformed manual annotations as well as predicted sentiment labels into numbers (-1 for the negative, 0 for the neutral and 1 for the positive class). In case of lexicon-based approaches we used a compound score per sentence instead, which already leads to values in the interval [-1, 1]. Afterwards, we calculated mean sentiment values per text and compared the values from our manual annotations to model predictions.

3.4 Explainability Analysis

In order to reach a deeper understanding of the main differences between the systems under investigation, we conducted various SHAP experiments. SHAP values are gained by masking an input as a whole before subsequently unmasking tokens. Thus, they measure the impact of a given token on the probability that the model under investigation predicts a given label on a range from -1 (negative impact) to 1 (positive impact) (Lundberg and Lee, 2017). It should also be noted that the sum of all SHAP values per class (three in our case) is always zero. Annotating all of our 442 test sentences this way took approximately one hour on a T4 GPU. We analyzed the enriched dataset by first looking at distributions per class. Afterwards, we calculated means per token, bigram and trigram and looked at the most impactful ones per class.

4 Results

We performed two types of model evaluations, namely looking at raw performance scores on the one hand and explainability attempts on the other. The following subsections present our findings.

[3] https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student

[4] https://github.com/tyqiangz/multilingual-sentiment-datasets

Model	Acc	F1 Scores		
		neg	neut	pos
Random Baseline	.33	.33	.33	.33
Majority Baseline	.40	.00	.00	.57
SentiWS	.34	.03	.51	.00
GerVADER	.59	.46	.59	.65
ger-senti-bert	.37	.18	.51	.09
senti-distilbert	.45	.56	.00	.53
bert*	.63	.54	.72	.63
gbert*	**.81**	**.76**	**.84**	**.82**
ger-senti-bert*	.74	.69	.77	.75

Table 2: Sentiment classification results. Models marked with * are fine-tuned on Moravian data, best results are highlighted in bold.

4.1 Sentiment Classification

Table 2 lists accuracy and F1 scores per class for each model based on predictions on the test dataset as well as random and majority baselines. In order to compare the different approaches, the compound scores of lexicon-based approaches (`SentiWS` and `GerVADER`) were categorized with thresholds of ±.05 that are said to yield best results (Hutto and Gilbert, 2014).

It is worth noting that the addition of context-aware rules, which is the main difference between the two lexicon-based approaches led to an .25 increase in accuracy to .59. The individual F1 scores imply that this may be due to a better recognition of non-neutral instances. Similar issues can be seen in the scores of `ger-senti-bert` explaining why this transformer-based approach is also outperformed by `GerVADER`. The multilingual `senti-distilbert` by contrast hardly identifies any neutral samples at all which is another notable finding since this is the only model doing so. The scores can be improved by fine-tuning as we showed with `ger-senti-bert`. This model ranks between base cased BERT models which is in line with previous research (Schmidt et al., 2021). In our case `gbert` performed best with an accuracy of .81 and similar F1 scores. Looking at individual F1 scores, all fine-tuned models share the common trait of performing worst in recognizing negative samples. This may be due to the fact that this class is underrepresented in our dataset (see Table 1).

Figure 2 shows confusion matrices of our fine-tuned models which illustrate that neither one of them had problems distinguishing neutral from non-neutral samples. However, the English BERT model confused about one third of the negative samples with positive ones and vice versa. A similar but weaker trend is also observable for the other fine-tuned models.

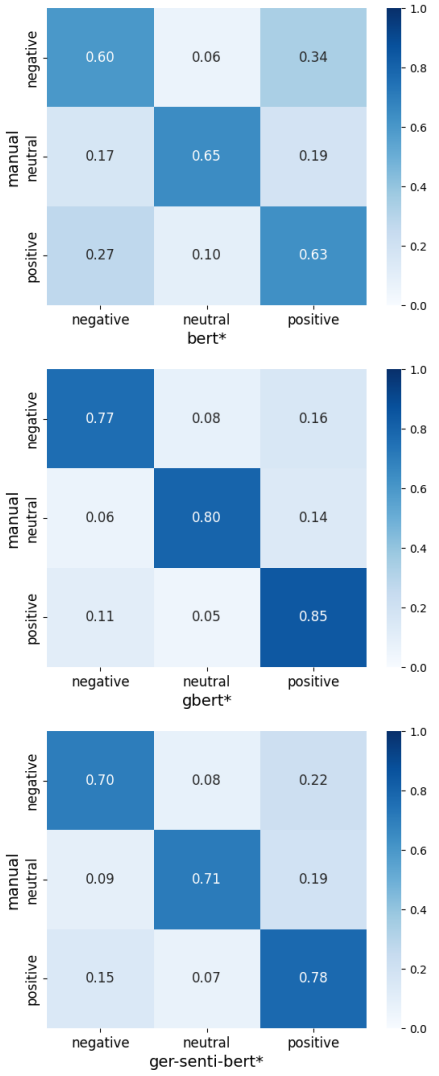

Figure 2: Confusion matrices of fine-tuned models

4.2 Error Analysis

Since the fine-tuned `gbert` model performed best, we focus our error analysis on this model. Even though many misclassifications cannot be categorized (see Table 3), we note that in a few cases, the model fails to consider the outcome in sentences with mixed sentiment. Example (1) from above is classified as `positive` but annotated as `negative`.

The most error-prone group were long sentences since they tend to contain mixed sentiment. Short sentences, by contrast, were classified wrongly in

Error	Confusion			
	neg/pos	neg/neut	neut/pos	**Total**
mixed	.08	.01	.01	.11
long	**.20**	.13	.02	.36
short	.01	.01	.07	.10
negated	.05	.04	–	.08
other	.10	**.14**	**.24**	**.48**
Total	**.44**	.33	.35	1.00

Table 3: Distribution of error types of the fine-tuned gbert model. Highest values are highlighted in bold.

a few cases where they were ambiguous without context. Finally, some errors can be explained by multiple negations as in (2).

(2) er würde nicht flüchten den er hätte den Indianern nichts böses sondern vielmehr gutes gethan,
('he would not flee because he had done the Native Americans no harm but rather good,')

4.3 Explaining Sentiment Predictions

Here, we present results from our explainability experiments. We start by using standard plotting functionality of the shap library[5] on a single prediction before analyzing SHAP value distributions and n-grams from our whole test dataset.

4.3.1 Explaining Individual Predictions

As SHAP is at its core a local explainability method, it offers ways to visualize which input features contributed to what degree to a model output. One such visualization can be seen in Figure 3 where (3) is analyzed.

(3) Die lezte Zeit kränckelte er.
('In recent times, he has been ailing.')

In this case, our fine-tuned English BERT model falsely predicts the positive class with 57.8 % confidence while the true label `negative` only reaches 34.7 %. The German model, by contrast, labels the sentence correctly with 99.8 % confidence. This difference between the two models is a typical one when considering their confusion matrices (see Figure 2) and can be explained in this specific instance. The former model focused on the wrong word forms namely *Die lezte* 'In recent' while the latter identified the correct sentiment anchor *kränckelte* 'ailing'. This can be seen by higher supporting SHAP values (colored in deeper red in Figure 3) attributed to the respective BERT tokens.

4.3.2 SHAP Value Distributions

To get insights into the general classification behavior of our fine-tuned models, we looked at the whole distribution of SHAP values (see Figure 4). As expected, the vast majority of individual tokens in our test dataset are non-discriminatory in nature. The means per class fall into the interval (-.01, .00) for the English BERT model and (-.01, .01) for the German one. Interestingly, the slightly negative mean SHAP value belongs to the neutral class in both cases. This trend is even more apparent when looking at outliers as the neutral class is the only one with considerably more negative outliers than positive ones. Thus, both models seem to recognize neutral samples stronger ex negativo which can be seen as a learning success since we used this class only in cases that lack sentiment bearing words.

Another interesting finding lies in the fact that fine-tuning `gbert` leads to considerably larger SHAP value intervals in all three classes. For example, the smallest range of this model was .94 with values between -.32 and .62 in case of attributions to the negative class. However, this is almost twice the maximum range of the fine-tuned `bert` (.56) that it reached with attributions to the neutral class with values between -.32 and .24. To sum it up, this implies that the German model was better at learning discriminatory tokens which matches our observations from looking at confusion matrices (see Figure 2) and will be investigated further in the following section.

4.3.3 N-gram Analysis

The additive nature of SHAP values enabled us to also look at token combinations (n-grams) that had the highest impact on predictions of our fine-tuned models. In Figure 5, we present SHAP values for single tokens and bigrams. Bigrams were added in order to get more interpretable results in our setting. This is due to the fact that especially tokens from `bert` were too ambiguous without context while an inclusion of trigrams led to a mere increase in variations of top ranking bigrams.

The English BERT has seemingly recognized the German negator *nicht* 'not' split into *ni* and *cht* as one of the most typical features of negative samples. This result was rather unexpected but gave

[5] https://shap.readthedocs.io/en/latest/

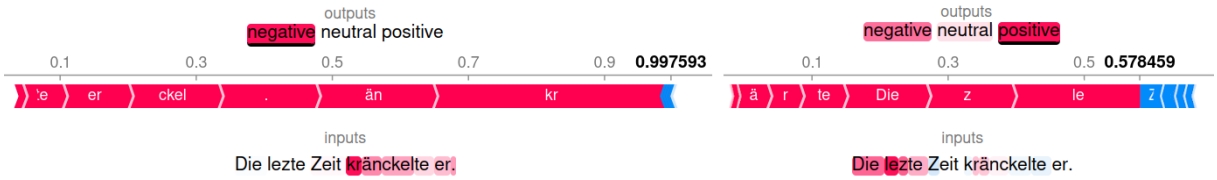

Figure 3: Default SHAP explanation plots per fine-tuned model (gbert* (left), bert* (right)). The predicted class is underlined and the confidence highlighted in bold. Negative SHAP values for the predicted class are colored blue and positive ones red. Lighter colors correspond to SHAP values closer to 0. The input sentence is shown in (3).

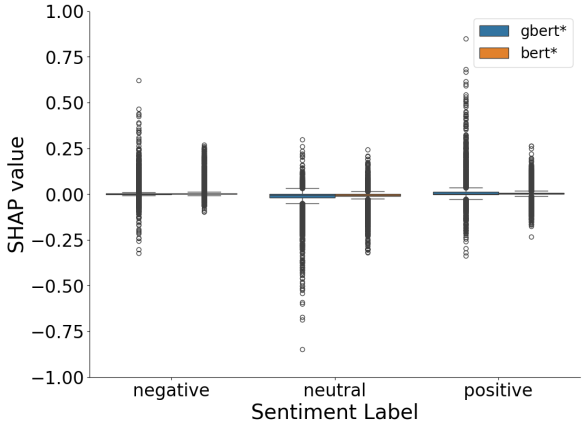

Figure 4: SHAP value distributions per model

insights into an imbalance in our train dataset as 29.5 % of the negative samples contain this type of negation but only 9.1 % percent of non-negative ones. Not only did gbert learn this feature but also some collocations of German words that bear negative sentiment like *Schmerz* 'pain', *verloren* 'lost', *Versuch[ung]* 'temptation', *Furcht* 'fear', *schwach* 'weak', *kon[fus]* 'confused', *schlecht* 'bad', *krank* 'sick' and *fehlen* 'to lack'.

With regard to neutral samples, it is worth mentioning that the English model considered dates and places as typical instances of this class. This is a learning success as such entities are commonly used in travel descriptions or introductory passages that tend to lack sentiment in Moravian memoirs. Interestingly, it also found the verb *heiraten* 'to marry' which accounts for the fact that marriages were indeed presented factually and not especially emotionalized in most memoirs. Top n-grams that contributed to neutral labeling of gbert on the other hand are harder to interpret. There is a prevalence of forms of the verb *wollen* 'to want' although they were similarly frequent in neutral annotations as in other ones.

The results for the positive class mirror those for the negative one. The English BERT seemed to have found another slight imbalance in our training corpus namely that exclamation marks occurred a little more frequently in positive samples (5.4 %) than in others (1.8 %). It might have also recognized the positively connoted noun *Herz* 'heart' as the trigram *zu +Her+zen* (.29) (c.f. *zu +Her* (.17) in Figure 5) implies. Still, gbert has once again learned more sentiment bearing lemmata like *lieben* 'to love', *gut* 'good', *willig* 'willing', *Vergnügen* 'pleasure', *helfen* 'to help', *Dank* 'gratitude', *Freude* 'joy' and *genießen* 'enjoy'.

Finally, we want to stress the fact that the mean SHAP values of n-grams from the German BERT model were consistently higher than ones from the English model even though is is less apparent for the neutral class. This is in line with our findings in section 4.3.1.

5 Applications

To illustrate some of the analyses made possible through sentiment assignments of the Moravian memoirs, we showcase two content-wise analyses of interest to the Moravian community.

5.1 Gender-based Sentiment Differences

We compared the German corpus with an English equivalent that was already analyzed (Faull and McGuire, 2022) by grouping sentiment annotations per gender. However, we did not limit our analysis to means but looked at whole distributions instead. Figure 6 shows the results with regard to manual annotations as well as ones generated from some of the models listed above. Note that we ignored the random and majority baselines since they are not expected to provide meaningful results. We also did not include SentiWS and our fine-tuned ger-senti-bert as they did not lead to insights that cannot be drawn from other models. This is due to the fact that the former classified almost anything as neutral and the latter was consistently in between the other two fine-tuned models.

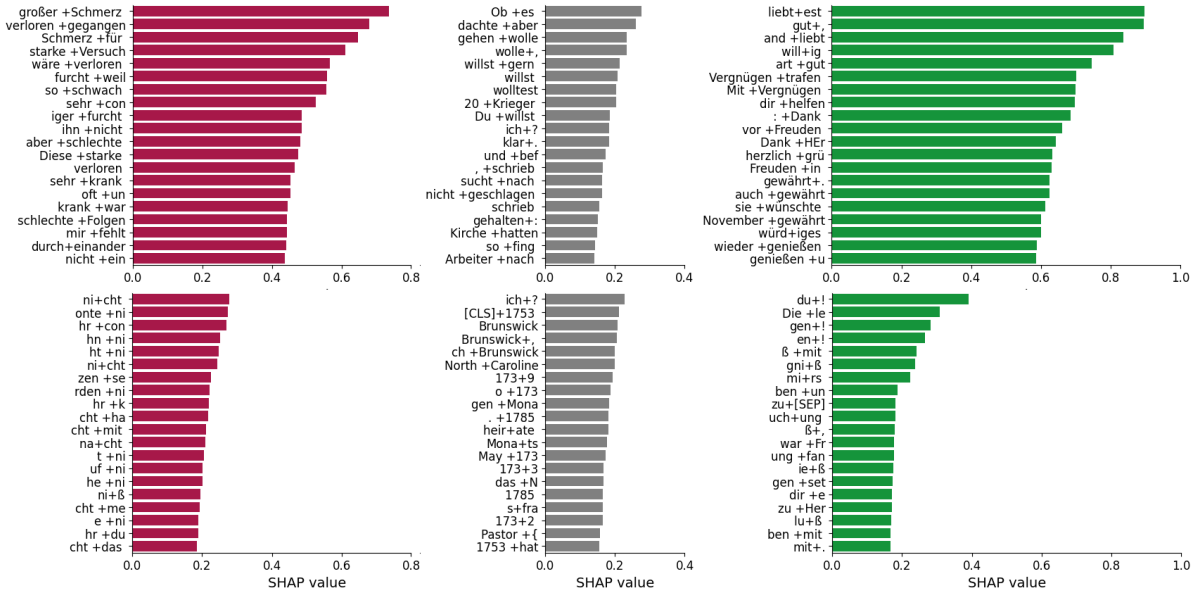

Figure 5: Single tokens and bigrams with highest SHAP values per class (negative (left), neutral (middle), positive (right)) and fine-tuned model (gbert* (top), bert* (bottom))

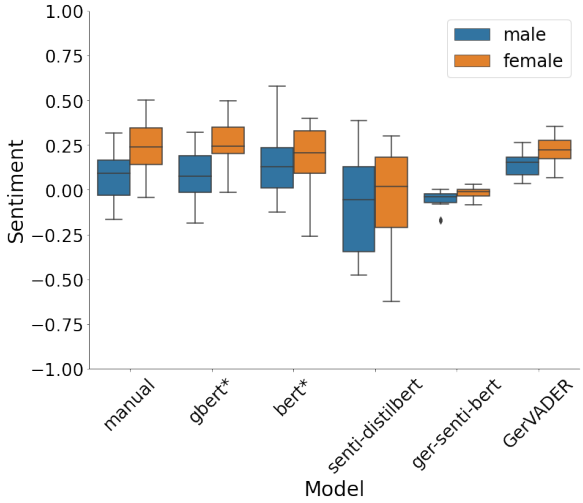

Figure 6: Sentiment distributions per gender and model

Our manual annotations are in line with the tendency found in the English corpus that lives of Moravian women are presented more positively than the ones of men (Faull and McGuire, 2022), especially when considering mean sentiment scores only. The fine-tuned `gbert` closely mirrored this distribution which was expected as it performed best (see Table 2). The same trend can however also be induced from most of the worse performing models although less overt. This is particularly true for the multilingual `senti-distilbert` which not only yields the highest range of sentiment values in general but also the biggest overlap between both distributions. On top of that, it leads to whiskers that are contrary to the general trend. Most of these findings are also true for classifications from the fine-tuned `bert`, albeit to a lesser extent. On the other hand, it is noteworthy that the sentiment values gained from `senti-distilbert` are more neutral (and negative). The former can also be seen by looking at the results of the other non-fine-tuned transformer model `ger-senti-bert` even though that model fails to identify most non-neutral samples (see Table 2). Interestingly, the lexicon-and-rule-based approach `GerVADER` performed almost as well on this task as our best fine-tuned model although its accuracy and F1 scores are worse.

5.2 Sentiment of "Blood and Wounds" Theology Related Words

Another research driven application is a quantitative analysis of the effects of "Blood and Wounds" theology on this corpus. The hypothesis is that tokens associated with this theology and accompanying themes are used in a positive context in memoirs of this specific time frame (Atwood, 2006; McGuire, 2021). Note that this was to the best of our knowledge not yet researched empirically, though. Nevertheless, most of the models we tested show this tendency as Table 4 illustrates.

Here, all fine-tuned models are able to confirm the hypothesis with mean sentiment values above .50. This suggests a strong positive sentiment towards the (sub)strings *[Bb]lut* 'blood'/'bleed' and *Wunden* 'wounds' which is in line with our man-

Model	Mean Sentiment	
	B&W	Dataset
manual	.52	.15
gbert*	**.52**	**.16**
bert*	.57	**.14**
senti-distilbert	.08	-.06
ger-senti-bert	-.01	-.04
GerVADER	.20	.18

Table 4: Mean sentiment per model of sentences with "Blood and Wounds" words (B&W) compared with the entire dataset. Models marked with * are fine-tuned on Moravian data, best results are highlighted in bold.

ual annotations. Our fine-tuned basic cased `bert` seemed to have indeed learned this very specific framing in Moravian memoirs of that time. It even slightly overrates the sentiment in relevant samples. It has to be noted, though, that both fine-tuned models have seen most of the samples as part of the train dataset. The other models we tested, on the other hand, hardly capture the fact that these tokens are as positively framed. This can be seen for example in the case of the weak positive sentiment attributed by `senti-distilbert`. This result is in line with its general performance on our data and the one on the previous task as the aggregation of mainly non-neutral annotations may lead to a mean value close to zero. `ger-senti-bert`, on the other hand, yields a very weak negative sentiment which can not only be explained in regard to its general classification tendency but also by the expected framing in non-Moravian data. Finally, `GerVADER` mirrors once again the general tendency but with a mean sentiment of .20 to a lesser extent.

6 Conclusion

In this paper, we introduced a manually annotated dataset for ternary sentiment analysis of German memoirs of Moravians that lived in the 18th century. The prevalence of non-neutral samples in it attest that sentiment in particular and emotions in general are important features of this domain which is in line with theological research (Van Gent, 2017; Faull and McGuire, 2022).

We also introduced BERT models fine-tuned on this dataset that not only outperform existing transformer models and lexicon-based approaches but also reach or even surpass state-of-the-art results (Allaith et al., 2023; Schmidt et al., 2021). This was not only true for performance statistics like accuracy and F1 scores but also for two research driven applications. Here, we found that German memoirs of women tend to be more positive than those of men and that a positively framed "Blood and Wounds" theology can be observed empirically. Both findings confirm results from various Moravian research projects (Atwood, 2006; Faull and McGuire, 2022; McGuire, 2021). They also show that the minimum performance level required may depend on the downstream task at hand.

Concerning model explainability, we showed that a deeper look at F1 score distributions and confusion matrices can already give some hints on the classification behavior and possible problems related to this. These results can be enriched by applying local explanation approaches like SHAP to a whole dataset. This revealed in our case that `gbert` actually learned sentiment bearing lemmata during fine-tuning and that a neutral class has to be inferred ex negativo. The base English BERT model, by contrast, focused more on rather random imbalances in our test dataset like negations and punctuation marks. The latter was also observed in a related NLP task (Inácio et al., 2023). We suspect that these differences might be due to the different tokenizers involved as the German one tends to split fewer word forms that may carry sentiment information.

From these observations we draw the conclusion that state-of-the-art performances for ternary sentiment analysis can already be reached with less than 2k fine-tuning samples. This makes transformer-based approaches feasible in low resource settings even more so since individual model predictions can be explained to a certain degree.

Limitations

This work should be seen as a case study that complements others like in the case of our performance comparisons (see Table 2) which confirmed trends from similar research projects. However, they are still only valid for our specific setting and thus we expect our fine-tuned models to perform worse when applied to strongly deviating domains. With regard to our explainability analysis, we want to stress the fact that the calculation of SHAP values is a resource-intensive task as also noted in another Sentiment Analysis application (Zielinski et al., 2023). In our case it was even more intensive than

the main fine-tuning step. This could be accounted for in a future ablation study.

Ethics Statement

It should be noted that our fine-tuned models might have learned derogatory language in relation to Native Americans or immigrants with non-European origins which should be seen in the given historical context. We refrained from masking corresponding terms in order to enable future research especially since other projects already deal with framing and devaluation phenomena in Moravian missionary narratives (Lasch, 2023).

References

Ali Allaith, Kirstine Degn, Alexander Conroy, Bolette S. Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment Classification of Historical Danish and Norwegian Literary Texts. In *Nordic Conference of Computational Linguistics*, pages 324–334. University of Tartu Library.

Craig D. Atwood. 2006. Understanding Zinzendorf's Blood and Wounds Theology. *Journal of Moravian History*, 1:31–47.

Luca Bacco, Andrea Cimino, Felice Dell'Orletta, and Mario Merone. 2021. Explainable Sentiment Analysis: A Hierarchical Transformer-Based Extractive Summarization Approach. *Electronics*, 10(18).

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proceedings of ACL-08: HLT*, pages 807–815, Columbus, Ohio. Association for Computational Linguistics.

Peter Burke. 2013. The rhetoric of autobiography in the seventeenth century. In Marijke J. van der Wal and Gijsbert Rutten, editors, *Touching the Past. Studies in the historical sociolinguistics of ego-documents*. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 633–642, Valencia, Spain. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Ashley Dennis-Henderson, Matthew Roughan, Lewis Mitchell, and Jonathan Tuke. 2020. Life still goes on: Analysing Australian WW1 diaries through distant reading. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 90–104, Online. International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esther Farbstein. 1998. Diaries and Memoirs as a Historical Source - The Diary and Memoir of a Rabbi at the "Konin House of Bondage". *Yad Vashem Studies*, XXVI:87–128.

Katherine Faull. 2021. *Visualizing religious networks, movements, and communities: building Moravian Lives*, pages 213–236. De Gruyter, Berlin, Boston.

Katherine Mary Faull and Michael A. McGuire. 2022. Analyzing Moravian Feelings Using Computational Methods to Ask Questions about Norms and Sentiments in Eighteenth-Century Moravian Lebensläufe. *Journal of Moravian History*, 22:125–149.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber. 2014. BiographyNet: Methodological Issues when NLP supports historical research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3728–3735, Reykjavik, Iceland. European Language Resources Association (ELRA).

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1620–1625, Marseille, France. European Language Resources Association.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605,

Austin, Texas. Association for Computational Linguistics.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Eero Hyvonen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 2019. Linked Data – A Paradigm Shift for Publishing and Using Biography Collections on the Semantic Web. In *Proceedings of the Third Conference on Biographical Data in a Digital World*, pages 16–23, Varna, Bulgaria.

Marcio Inácio, Gabriela Wick-Pedro, and Hugo Goncalo Oliveira. 2023. What do Humor Classifiers Learn? An Attempt to Explain Humor Recognition Models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Hiroaki Kuromiya. 1985. Soviet Memoirs As A Historical Source. *Russian History*, 12:293–326.

Alexander Lasch. 2023. Unterschiede „zwischen uns & den weißen Leuten“. In *Die Herrnhuter Brüdergemeine im 18. und 19. Jahrhundert*, volume 69 of *Arbeiten zur Geschichte des Pietismus*, pages 531–550. Vandenhoeck & Ruprecht.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris B. Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michael McGuire. 2021. *Computational Sentiment Analysis of an 18th Century Corpus of Moravian English Memoirs*. Ph.D. thesis, Indiana University.

Geoff Mortimer. 2002. *Eyewitness Accounts of the Thirty Years War 1618–48*. Palgrave Macmillan.

Eric T. Nalisnick and Henry S. Baird. 2013. Character-to-Character Sentiment Analysis in Shakespeare's Plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, Sofia, Bulgaria. Association for Computational Linguistics.

Alessio Palmero Aprosio and Sara Tonelli. 2015. Recognizing biographical sections in Wikipedia. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Lisbon, Portugal. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Fritz Redlich. 1975. Autobiographies as sources for social history: A research program. *VSWG: Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte*, 62(3):380–390.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS: A Publicly Available German-language Resource for Sentiment Analysis. In *International Conference on Language Resources and Evaluation*, pages 1168–1171.

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

Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. 2019. GerVADER: A German Adaptation of the VADER Sentiment Analysis Tool for Social Media Texts. In *Lernen, Wissen, Daten, Analysen*.

Jacqueline Van Gent. 2017. Moravian Memoirs and the Emotional Salience of Conversion Rituals. In *Emotion, Ritual and Power in Europe, 1200–1920: Family, State and Church*, pages 241–260.

Florian Windhager, Matthias Schlögl, Maximilian Kaiser, Ágoston Zénó Bernád, Saminu Salisu, and Eva Mayr. 2017. Beyond One-Dimensional Portraits: A Synoptic Approach to the Visual Analysis of Biography Data. In *Proceedings of the Second Conference on Biographical Data in a Digital World*, pages 67–75, Linz, Austria.

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. Prediction of Happy Endings in German Novels based on Sentiment Information. In *Proceedings of DMNLP, Workshop at ECML/PKDD*, pages 9–16, Riva del Garda, Italy.

Andrea Zielinski, Calvin Spolwind, Henning Kroll, and Anna Grimm. 2023. A Dataset for Explainable Sentiment Analysis in the German Automotive Industry. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 138–148, Toronto, Canada. Association for Computational Linguistics.