Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language

Thomas Schmidt
Media Informatics Group
University of Regensburg, Germany
thomas.schmidt@ur.de

Katrin Dennerlein
German Literary Studies
University of Würzburg, Germany
katrin.dennerlein
@uni-wuerzburg.de

Christian Wolff
Media Informatics Group
University of Regensburg, Germany
christian.wolff@ur.de

Abstract

We present results of a project on emotion classification on historical German plays of *Enlightenment*, *Storm and Stress*, and *German Classicism*. We have developed a hierarchical annotation scheme consisting of 13 sub-emotions like *suffering*, *love* and *joy* that sum up to 6 main and 2 polarity classes (positive/negative). We have conducted textual annotations on 11 German plays and have acquired over 13,000 emotion annotations by two annotators per play. We have evaluated multiple traditional machine learning approaches as well as transformer-based models pretrained on historical and contemporary language for a single-label text sequence emotion classification for the different emotion categories. The evaluation is carried out on three different instances of the corpus: (1) taking all annotations, (2) filtering overlapping annotations by annotators, (3) applying a heuristic for speech-based analysis. Best results are achieved on the filtered corpus with the best models being large transformer-based models pretrained on contemporary German language. For the polarity classification accuracies of up to 90% are achieved. The accuracies become lower for settings with a higher number of classes, achieving 66% for 13 sub-emotions. Further pretraining of a historical model with a corpus of dramatic texts led to no improvements.

1 Introduction

Transformer-based language models like BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2019) have recently gained a lot of attention and achieve state-of-the-art results for various tasks in natural language processing (NLP) (Qiu et al., 2020). These language models are usually trained via deep learning on large amounts of texts acquired from the internet. Unlike previous methods in NLP, these models use context-sensitive word representations and they can better deal with out-of-vocabulary words. These attributes are, of course, advantageous for various text sorts in digital humanities (DH) and computational literary studies (CLS). Furthermore, transformer-based language models can be adapted to specific domain texts by either training a model from scratch on large amounts of these texts or taking an existing model and further pretraining it with domain-specific texts (Beltagy et al., 2019; Gururangan et al., 2020; Rietzler et al., 2020). Transformer-based models as well as these approaches have been successfully applied in DH contexts with historical or poetic German texts for named entity recognition (NER) (Schweter and Baiter, 2019; Labusch et al., 2019) and speech type recognition (Brunner et al., 2020). We present a study for the task of textual emotion classification in the same line of research for the use case of German historical plays.

Emotion classification deals with the prediction of (multiple) emotion categories in text. Its neighbouring field sentiment analysis primarily focuses on the prediction of the overall polarity (or valence) of a text, meaning if it is rather positive or negative (Mäntylä et al., 2018). Both methods have been explored in DH and CLS to analyze emotion/sentiment distributions and progressions in social media (Schmidt et al., 2020b) or literary texts like plays (Nalisnick and Baird, 2013; Schmidt and Burghardt, 2018; Schmidt et al., 2019b; Schmidt, 2019), novels (Zehe et al., 2016; Reagan et al., 2016) and fairy tales (Alm and Sproat, 2005; Mo-

Proceedings of LaTeCH-CLfL 2021, pages 67–79
Punta Cana, Dominican Republic (Online), November 11, 2021.

hammad, 2011) (see Kim and Klinger (2019) for an in-depth review of this research area). However, as the review of Kim and Klinger (2019) and recent tool developments in DH show (Schmidt et al., 2021a), the application of rather basic lexicon-based methods is frequent although these methods are usually outperformed by more modern approaches in sentiment and emotion classification (Cao et al., 2020; Dang et al., 2020; Cortiz, 2021; González-Carvajal and Garrido-Merchán, 2021) and are especially problematic for literary texts (Fehle et al., 2021). Furthermore, performance evaluation of computational approaches compared to human annotations ("gold standard") are rare. Thus, we present an evaluation study for the use case of German historical plays (from *Enlightenment*, *Storm and Stress* and *German Classicism*) for emotion classification. Our goal is to develop emotion classification algorithms with a satisfactory performance for the described use case to investigate in later stages of our research, for example, emotion progressions throughout time or genre-based differences concerning emotion distributions on a larger set of plays. We primarily focus on current state-of-the-art transformer-based language models.

The main contributions of this paper are as follows: (1) the development of an emotion annotation scheme directed towards the interest of literary scholars for the time frame of our corpus, (2) the annotation results for the annotation of 11 plays by 2 annotators for each play, (3) a systematic evaluation of traditional textual machine learning (ML)-approaches, transformer-based models pretrained on contemporary and historical language and further pretrained on dramatic texts on different instances of the annotated corpus. The goal of this contribution is to work towards the development of emotion classification algorithms with a satisfactory performance for the described use case.

2 Training and Evaluation Corpus

In the following section, we describe the conceptual framework and process for the acquisition of the annotated corpus that serves as training and evaluation corpus for the emotion classification ("Gold Standard").

2.1 Emotion Scheme

The main goal of the scheme development was to create an annotation scheme that includes the interests of literary scholars and the interpretative and historical dimensions of these literary texts. Common emotion annotation schemes in NLP are mostly inspired by psychology, oftentimes consisting of 6-8 established emotion classes (cf. Wood et al., 2018a,b). However, we regard these concept sets as unfit for our specific case, since important emotion and affect concepts from the perspective of literary criticism for the time of our plays are missing, while other concepts are not specifically important for our text genre. Thus, we developed a novel annotation scheme based on literary theory and re-designed the scheme in an iterative process of small pilot annotations and discussions. Our final scheme deviates heavily from more common schemes in emotion annotation in NLP. Some concepts well-known in NLP and psychology are included like *joy*, *fear* or *anger* while other standard emotion concepts like *disgust* and *surprise* showed in pilot annotations to be not of great importance. Concepts, important for literary critique for that time that are not usually regarded as emotions, that we include are *desire*, *suffering* or *compassion*. Please refer to Schmidt et al., (2021b) for more information about the scheme creation and the annotation process.

The final scheme consists of 13 *sub-emotions* that are hierarchically clustered including one special concept (*emotional movement*). The *sub-emotions* are summarized in six *main classes* which can then be clustered in a final binary setting of two classes (similar to sentiment): (per default) positive and negative emotions (marked in the upcoming list as + and - respectively; we refer to this concept as *polarity*). In the following list we name the *sub-emotions* and *main classes* with the original German term in brackets (since we perform annotations in German) and an English translation.

- emotions of affection (*Emotionen der Zuneigung*)
  - desire *(Lust)* (+)
  - love *(Liebe)* (+)
  - friendship *(Freundschaft)* (+)
  - admiration *(Verehrung)* (+)
- emotions of joy *(Emotionen der Freude)*
  - joy *(Freude)* (+)

  - Schadenfreude (The joy about the misfortune of others) (+)

- emotions of fear *(Emotionen der Angst)*
  - fear *(Angst)* (-)
  - despair *(Verzweiflung)* (-)

- emotions of rejection *(Emotionen der Ablehnung)*
  - anger *(Ärger)* (-)
  - hate, disgust *(Hass, Abscheu)* (-)

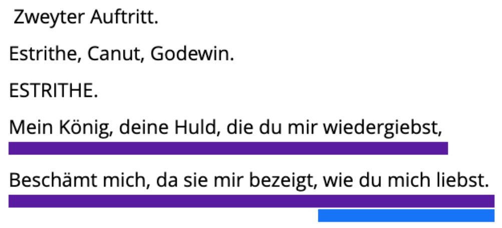
- emotions of suffering *(Emotionen des Leids)*
  - suffering *(Leid)* (-)
  - compassion *(Mitleid)* (-)

- emotional movement

Emotional movement has no polarity and is used to describe astonishment, emotional turmoil, excitation and oscillation between several emotions. We will refer to the combination of the *positive*- and the *negative*-class as well as *emotional movement* as *triple polarity*. The various hierarchical structures are later used for classification approaches with different class numbers.

2.2 Annotation Process

Annotators are instructed to assign sub-emotions, as defined in our scheme, to text. We regard the character's state of mind as expressed in the text as the emotion to be annotated. Annotations are performed context-sensitive, meaning annotators should take into account the plot and content of the entire play and annotate what the character really means as determined by the literary interpretation. Thus, plays are read and annotated from beginning to end concerning *stage directions* and *speeches* (single utterances of characters separated by the utterances of other characters). Depending on the emotional expression in the text, annotators can mark text sequences of varied lengths (ranging from one word to an entire speech) and are not limited to a concrete annotation size. Furthermore, annotators can annotate multiple annotations per text sequence fully or partially (see figure 1) and adjust the default polarity of sub-emotions for certain cases. The annotation procedure just presented is closer to the interpretation process of literary scholars than context-free approaches with fixed text sizes for annotation attribution that are more common in NLP (Mäntylä et al., 2018). It has been deemed as more fitting throughout multiple pilot annotations with literary scholars.



Zweyter Auftritt.

Estrithe, Canut, Godewin.

ESTRITHE.

Mein König, deine Huld, die du mir wiedergiebst,

Beschämt mich, da sie mir bezeigt, wie du mich liebst.

Figure 1: Example annotation in CATMA. The annotator marked two lines as *suffering* (purple), and the last part additionally as *love* (blue). (Excerpt from *Canut*)

The annotation process itself is performed with the tool CATMA[1] (Gius et al., 2020). Two annotators annotate each play independently from each other in a time span of 1-2 weeks depending on the length of the play. All annotators are students of German literary studies and are compensated monetarily for the annotation. They have access to an annotation instruction manual with descriptions of the scheme and examples. They also participated in test annotations under the guidance of an expert literary scholar. Indeed, the entire annotation process is iterative (cf. Reiter, 2020) meaning scheme and instructions changed based on feedback throughout the project cycle and might be due to change (the study presented here has been performed consistently in the way described, however).

2.3 Annotated Plays

As part of our larger project, we intend to analyze emotion classification on historical German plays between 1650-1815. Our current corpus of plays consists of around 300 digitized plays. For this evaluation study, we annotated a representative sub-corpus of 11 plays of varying epochs and genres. However, we focus on more recent plays for this first evaluation study since older ones are more likely to pose more challenges to the applied language models:

- *Das Testament* by Gottsched (1745/comedy)

- *Canut* by Schlegel (1746/tragedy)

- *Die zärtlichen Schwestern* by Gellert (1747/comedy)

- *Lucie Woodvil* by Pfeil (1757/tragedy)

[1] https://catma.de/

- *Der Freigeist* by Brawe (1758/tragedy)
- *Minna von Barnhelm* by Lessing (1767/comedy)
- *Der Postzug* by Ayrenhoff (1769/comedy)
- *Kabale und Liebe* by Schiller (1784/tragedy)
- *Kasperl' der Mandolettikrämer* by Eberl (1789/tragedy)
- *Menschenhass und Reue* by Kotzebue (1790/comedy)
- *Faust* by Goethe (1807/tragedy)

Most of the plays were acquired as part of the *GerDracor*-Corpus (Fischer et al., 2019) except for *Kasperl' der Mandolettikrämer* which was acquired via an open web repository.[2]

2.4 Annotation Statistics

Depending on the length of a play, the annotation duration for each play was 8-15 hours in absolute numbers. We collected 13,264 annotations of varying lengths. Table 1 illustrates the distributions for the *sub-emotions* as well as the resulting sums for the *main classes*.

The most frequent *sub-emotion* are *suffering* (16%) and *love* (13%) and the *emotions of rejection* and (23%) *affection* (22%) for the *main classes* respectively. The overall distribution is rather imbalanced with certain sub-emotions being rarely annotated (e.g. *desire*). Considering the overall *triple polarity*, the majority of annotations are *negative* (53%), followed by *positive* (37%) and *emotional movement* (11%). We also examined token statistics about annotation lengths: On average an annotation consists of 25 tokens, however with a large variance ranging from 1-token annotations to multiple sentences consisting of over 500 tokens. This shows that annotators make significant use of the possibility of varied annotation lengths.

Due to the varied annotation lengths, calculating inter-annotator agreement is not possible with common metrics. However, to get an overall understanding of the agreement we calculate agreement according to the following speech-based heuristic: For each speech, the emotion that is annotated the most per speech (measured in number of tokens) is assigned the specific emotion (or a neutral class if no emotion is annotated) for each annotator. This results in a *Cohen's* κ value of 0.5 for *polarity* (percentage wise agreement: 68%) and 0.4 for *main classes* (62%) and *sub-emotions* (58%) respectively. This is regarded as moderate agreement (Landis and Koch, 1977), which is low compared with sentiment analysis research with other text sorts (cf. Mäntylä et al., 2018) but in line with similar annotation projects with literary and historical texts (Alm and Sproat, 2005; Sprugnoli et al., 2015; Schmidt et al., 2018, 2019a,c; Öhman, 2020; Schmidt et al., 2020a).

Emotion category	absolute	%
MC: emotions of affection	2,928	22
desire	52	0
love	1,755	13
friendship	345	3
admiration	776	6
MC: emotions of joy	1,943	15
joy	1,619	12
Schadenfreude	324	2
MC: emotions of fear	1,257	9
fear	721	5
despair	536	4
MC: emotions of rejection	3,028	23
anger	1,625	12
hate, disgust	1,403	11
MC: emotions of suffering	2,700	20
suffering	2,069	16
compassion	631	5
emotional movement	1,408	11
Overall	13,264	100

Table 1: Distribution of emotion categories. First, the summed results of the *main classes* (MC; marked in bold) are listed followed by the *sub-emotions*. Percentages are rounded.

2.5 Corpus Manifestations

Due to the varied annotation text sequence lengths and the moderate agreement statistics, we evaluated and trained the chosen emotion classification approaches on different "manifestations" of our corpus. We refer to the first one as *full corpus*. This manifestation includes all text annotations of the two annotators for every play. Thus, it does include annotations for which the annotators disagree upon fully or partially. This is the largest corpus manifestation consisting of 13,264 annotations (for more statistics see table 1). For the classification of

[2] http://lithes.uni-graz.at/maezene/eberl_mandolettikraemer.html

polarity, we reduce corpora by filtering out annotations with *emotional movement*, which results in 11,883 annotations for the *full corpus*. The second manifestation is referred to as *filtered corpus*. For this corpus instance, we filter out all annotations for which annotators either fully or partially disagree, meaning annotations of different categories that overlap at least for one token. We do not filter out annotated text sequences by one annotator that are not annotated by the other one. We do however filter all overlapping contrary annotations by a single annotator. While our annotation scheme enables these kind of annotations, we want to evaluate how the filtering of all contrary overlaps influences emotion classification. Depending of the emotion hierarchy, this results in different annotation numbers for the final filtered corpus: 9,962 (*polarity*), 10,247 (*triple polarity*), 8,552 (*main class*), 7,503 (*sub-emotions*). Thus, the filtering reduces the corpus size between 15-44% depending of the categorical system.

The last manifestation, the *speech corpus*, is focused on the central units of plays: speeches and stage directions. It is designed as follows: Each speech (we include stage directions in the following when speaking about speeches) of the plays is assigned with the emotion category that is annotated the most by both annotators (as measured by number of tokens). If tied among multiple classes, the class is assigned that is overall chosen the least (to counteract class imbalances). The entire corpus consists of 11,617 speeches; we filter out speeches with no annotation by either annotator to avoid adding an extra neutral-like class to our already multi-class setting (adding neutrality is something we intend to explore in future work). This reduces the amount of speeches to 6,741 and affects especially stage directions which are rarely annotated. We apply the above heuristic to acquire emotion assignments. Please note that emotion distributions change compared to the other manifestations since underrepresented classes become even more rare due to the applied heuristic; thus the class imbalances intensify. Distribution statistics for the *filtered* and *speech corpus* can be found in the appendix (table 6, 7, 8).

We separate the corpus in these three manifestations in order to explore performance on different classification levels and text sizes, which will influence our decision for later large-scale emotion prediction tasks on larger corpora of plays which we plan for future stages of our project.

3 Emotion Classification Methods

We regard the emotion classification as single-label classification on text sequences of varied lengths. The amount of classes differs depending on the hierarchical system: *polarity* (2 classes), *triple valence* (3 classes), *main classes* (6), *sub-emotions* (13). We have implemented reference baselines based on traditional ML-approaches but otherwise focus on transformer-based language models for German pretrained on contemporary and historical texts since transformer-based models have been shown to achieve state-of-the-art results for emotion classification (Shmueli and Ku, 2019; Yang et al., 2019; Cao et al., 2020) and performed best in a pre-study (Schmidt et al., 2021c). We also explore further fine-tuning/pretraining of a pretrained model with our domain texts since research suggests performance improvements for this method (Beltagy et al., 2019; Gururangan et al., 2020; Rietzler et al., 2020).

3.1 Baseline Methods

The following "classical" ML-methods for text are implemented (methods like this are usually outperformed by transformer-based approaches in other settings (González-Carvajal and Garrido-Merchán, 2021) and thus serve as lower baselines in the following evaluation): (1) Representation of text units with term frequencies in a bag-of-words model and subsequently *Multinomial Naive Bayes* as training algorithm. (2) Same representation format as above but *Support Vector Machines* as training algorithm.

We implemented the approaches with the *scikit-learn machine learning library*[3] (Pedregosa et al., 2011) and trained and evaluated the algorithms in a stratified 5x5 cross evaluation setting. We refer to the first approach as *bow-nb* and the second one as *bow-svm*. We will also report the random and majority baseline for each classification task. Please note that depending on the corpus type, these values migh vary.

3.2 Transformer-based Models

We selected the (to our knowledge) most well-known and established transformer-based language models in German that are freely available. Table 2 summarizes the selected models (the identifiers

[3] https://scikit-learn.org/stable/

are used in the following to reference the models). All models are acquired via the *Hugging Face platform*[4] and are also implemented with the corresponding library (Wolf et al., 2020).

One main point of interest are performance differences between models pretrained on contemporary texts (e.g. like the *Wikipedia*, subtitles etc.) for general purpose tasks and models pretrained on historical texts (e.g. historical newspapers, historical fictional texts). In Table 2 we attribute the label "historical" to a model if a significant part of the texts dates from before the 20th century. We want to evaluate if these models perform better since the language is closer to the ones of our plays, which are of the 18th and 19th century.

For the contemporary models, we evaluate, among others, the models *gbert-large* and *gelectra-large* by *Deepset*[5] which achieve state-of-the-art results in standardized NLP-tasks (Chan et al., 2020) and are, to our knowledge, the largest German BERT- and ELECTRA-based models. On the historical side, we evaluate two models provided by the European Digital Library *Europeana* pretrained on historical newspaper (Schweter and Baiter, 2019; Schweter, 2020) and a model focused on fictional texts (Brunner et al., 2020). To perform the training and evaluation, each model is fine-tuned to the downstream task of emotion classification for the specific hierarchy and corpus. We apply the recommended settings for the training of downstream tasks, depending on the architecture: BERT (Devlin et al., 2019) or ELECTRA (Clark et al., 2019) as well as by the *Hugging Face*-library. Each model is fine-tuned for 4 epochs, a batch size of 32, learning rate of 4e-5 and *Adam* optimizer for stochastic gradient descent. The models are trained and evaluated in a 5x5 cross evaluation setting, thus averages over 5 runs are reported. As GPU a *Tesla P100* was used.

All of the above models are trained from scratch on large amounts of texts. However, recent research also suggests further pretraining of already existing models with texts that are close to the texts of the downstream task may improve results (domain-specific fine-tuning) (Gururangan et al., 2020; Rietzler et al., 2020). We explore this approach and further pretrain the model *bert-base-german-europeana-cased* solely with German dramatic texts that we acquired of our corpus sources (including the annotated texts). The texts consist of all German plays of *GerDracor* (Fischer et al., 2019), the platform *TextGrid*[6] and around 60 plays we acquired via various other sources. Altogether the texts sum up to 300 MB consisting of 1,224 plays that range from the 16th to the 20th century. We use the *simpletransformer*-library[7] and further pretrain the model *bert-base-german-europeana-cased* for 10 epochs. The setting and parameters for the emotion classification training are the same as for the general models. We refer to this model as *bert-europeana-further-pretrained*.

4 Results

We report accuracies and F1-scores for all models and category systems as well as corpus manifestations in tables 3, 4 and 5. Considering F1-scores, we report weighted F1 due to the imbalanced class distributions.

In general, transformer-based models outperform traditional ML-approaches. For every corpus manifestation the performance of the different transformer-based models is rather similar regardless whether contemporary or historical language is the basis for the pretraining. The best models are the large contemporary models *gbert-large* and *gelectra-large* achieving up to 90% for *polarity* (2 classes), 85% for *triple polarity* (3 classes), 75% for *main classes* (6 classes) and 66% *for sub-emotions* (13 classes) on the *filtered corpus*. The historical models perform rather similar but consistently slightly below the large contemporary ones, but also slightly above the smaller contemporary model *bert-base-german-europeana-cased*. Considering the different corpus manifestations, all models perform best on the *filtered corpus* and worst for the speech-based prediction. The difference becomes larger with increasing number of classes. For example, *gbert-large* achieves an accuracy of 75% for main class prediction on the *filtered corpus* which reduces to 51% on the *speech corpus*. As the analysis of recall and F1-macro statistics show, this is mostly due to the bad prediction accuracies for low-frequency classes.[8]

Further pretraining the model *bert-base-german-europeana-cased* with dramatic texts did not result

[4] https://huggingface.co/

[5] https://deepset.ai/german-bert

[6] https://textgrid.de/digitale-bibliothek

[7] https://simpletransformers.ai/

[8] Additional data about the results can be found via the following repository: https://github.com/lauchblatt/Emotions_in_Drama

Identifier	Hugging Face-identifier	Pretrained language	Pretrained texts and size if reported	Related paper (if available) and provider
bert-base	bert-base-german-cased	contemporary	Wikipedia, legal texts, news (∼12 GB)	Deepset
gbert-large	gbert-large	contemporary	Crawled web data, Wikipedia, subtitles, book, legal texts (∼161 GB)	Deepset (Chan et al., 2020)
gelectra-large	gelectra-large	contemporary	Crawled web data, Wikipedia, subtitles, book, legal texts (∼161 GB)	Deepset (Chan et al., 2020)
bert-europeana	bert-base-german-europeana-cased	historical	*Europeana* newspaper (51 GB)	MDZ Digital Library (Schweter, 2020)
electra-europeana	electra-base-german-europeana-cased-discriminator	historical	*Europeana* newspaper (51 GB)	MDZ Digital Library (Schweter, 2020)
bert-historical-rw	bert-base-historical-german-rw-cased	historical	Fairy tales, historical newspapers, magazine articles, narrative texts, texts of Projekt Gutenberg	(Brunner et al., 2020)
bert-europeana-further-pretrained	-	contemporary, further pretrained on historical texts	Based on *bert-base-german-europeana-cased*. Further pretrained with dramatic texts of *GerDracor*, *TextGrid* and other (300 MB)	-

Table 2: Transformer-based models for the evaluation. *Hugging Face*-identifier can be used to retrieve the models from the *Hugging Face*-platform, bert-europeana-further-pretrained was created by the authors of this paper via further pretraining.

in improvements. Indeed, the accuracies become slightly worse and significantly lower looking at settings with multiple classes (e.g. 29% for *sub-emotions* on the *filtered corpus*).

5 Discussion and Future Work

As the results show, we can confirm general findings of NLP-research for classification tasks for various text genres, in the sense that transformer-based models perform better than traditional textual ML-approaches in our setting with German historical plays. However, we cannot confirm our assumption that models pretrained on historical language achieve better results because they are closer to the language of our annotated material. Indeed, the best performing models are *gbert-large* and *gelectra-large* by *deepset* (Chan et al., 2020). These are, to our knowledge, the largest German models trained on contemporary texts, primarly internet texts. The difference between historical and these contemporary models is however small. Since the differences in the amount of text for the pretraining are significant (around 20 GB) it opens up the question if the performance of historical models improves with similarly large amounts of texts.

Considering the different corpus instances, we showed that filtering out overlapping annotations annotators disagree upon results into the strongest performance boost, although the training and test size become smaller. Thus, it is crucial for our project to find ways to deal with disagreements among annotators. Due to the varied and overlapping annotation lengths, we cannot rely on standard solutions like majority voting. Furthermore, the inherent subjectivity of literary texts and the resulting low agreement among annotators is a specific feature of these kind of texts. We do however think that we can reduce disagreement with further training of the annotators and also by implementing a subsequent step after the first annotations of two annotators, in which a literary scholar expert creates a *consensus annotation* resolving disagreement. Additionally, we intend to switch from single-label classification to multi-label emotion classification since this is more in line with the annotation process. This will open up further possibilities to deal with overlapping annotations and integrates this phenomenon into the classification task. Applying a heuristic to map single emotion classes to entire speeches led to the models performing rather poorly compared to the other corpus manifestations. For *sub-emotion* prediction with 13 classes, accuracies became 25% worse for certain

Method	**acc (pol)**	**F1 (pol)**	**acc (t-p)**	**F1 (t-p)**	**acc (m-c)**	**F1 (m-c)**	**acc (s-e)**	**F1 (s-e)**
random baseline	0.50	-	0.33	-	0.17	-	0.08	-
majority baseline	0.59	-	0.53	-	0.23	-	0.16	-
bow-svm	0.74	0.72	0.66	0.62	0.47	0.45	0.35	0.32
bow-bayes	0.78	0.78	0.70	0.68	0.52	0.50	0.39	0.35
bert-base	0.83	0.83	0.76	0.76	0.60	0.60	0.49	0.48
bert-europeana	0.84	0.84	0.76	0.76	0.61	0.61	0.50	0.49
electra-europeana	0.84	0.84	0.77	0.76	0.61	0.61	0.50	0.48
bert-historical-rw	0.84	0.84	0.76	0.76	0.61	0.61	0.51	0.50
gbert-large	**0.85**	0.85	**0.78**	0.77	0.63	0.63	0.52	0.52
gelectra-large	**0.85**	0.85	**0.78**	0.78	**0.64**	0.64	**0.53**	0.52
bert-europeana-further-pretrained	0.81	0.81	0.74	0.71	0.53	0.50	0.38	0.32

Table 3: Evaluation results for the *full corpus*. F1-scores are weighted F1. pol=*polarity*, t-p=*triple polarity*, m-c=*main class*, s-e=*sub-emotion*. Best result per classification is marked in bold for accuracies.

Method	**acc (pol)**	**F1 (pol)**	**acc (t-p)**	**F1 (t-p)**	**acc (m-c)**	**F1 (m-c)**	**acc (s-e)**	**F1 (s-e)**
random baseline	0.50	-	0.33	-	0.17	-	0.08	-
majority baseline	0.60	-	0.55	-	0.25	-	0.15	-
bow-svm	0.77	0.75	0.70	0.66	0.53	0.51	0.41	0.38
bow-bayes	0.83	0.83	0.76	0.74	0.59	0.56	0.46	0.41
bert-base	0.88	0.88	0.83	0.83	0.70	0.70	0.61	0.60
bert-europeana	0.88	0.88	0.83	0.83	0.71	0.70	0.60	0.59
electra-europeana´	0.89	0.89	0.83	0.83	0.70	0.69	0.56	0.53
bert-historical-rw	0.88	0.88	0.83	0.83	0.72	0.72	0.63	0.63
gbert-large	0.89	0.89	0.84	0.84	**0.75**	0.75	**0.66**	0.66
gelectra-large	**0.90**	0.90	**0.85**	0.85	0.74	0.74	0.64	0.63
bert-europeana-further-pretrained	0.83	0.83	0.76	0.74	0.45	0.38	0.29	0.23

Table 4: Evaluation results for the *filtered corpus*. F1-scores are weighted F1. pol=*polarity*, t-p=*triple polarity*, m-c=*main class*, s-e=*sub-emotion*. Best result per classification is marked in bold for accuracies.

models. While one reason is that this corpus is the smallest of all manifestations, we argue that the main problem is, as annotations showed, that most speeches consist of multiple, oftentimes differing emotion categories. Mapping them heuristically to one results in text units including various emotional expressions that are falsely mapped to one emotion. This problem intensifies due to the fact that many speeches are rather long and that the class imbalances for *main classes* and *sub-emotions* are significant. Thus, we plan to focus on smaller text unit sizes like sentences or n-grams in the future emotion prediction task over the entire corpus.

Considering the results for *filtered* and *full corpus*, the transformer-based models achieve state-of-the-art accuracies for polarity classification (88-90%) compared to results with sentiment analysis with similar amounts of classes on contemporary German (Chan et al., 2020). The results achieved by the transformer-based models for *polarity* are also around 20% above results on German dramatic texts predicted by lexicon-based sentiment analysis, which yields results around 70% (Schmidt and Burghardt, 2018). For the *main class* and *sub-emotion* classification, results are, however, for the best models (75% for main classes, 66% for sub-emotions), below state-of-the-art results on emotion classification tasks with 4 or more classes for contemporary English texts for which accuracies of up to 86% are reported (Shmueli and Ku,

Method	acc (pol)	F1 (pol)	acc (t-p)	F1 (t-p)	acc (m-c)	F1 (m-c)	acc (s-e)	F1 (s-e)
random baseline	0.50	-	0.33	-	0.17	-	0.08	-
majority baseline	0.60	-	0.51	-	0.23	-	0.16	-
bow-svm	0.62	0.53	0.53	0.42	0.28	0.23	0.22	0.16
bow-bayes	0.70	0.69	0.59	0.55	0.39	0.36	0.29	0.25
bert-base	0.73	0.73	0.63	0.62	0.46	0.45	0.36	0.34
bert-europeana	0.72	0.69	0.66	0.65	0.48	0.47	0.37	0.35
electra-europeana	0.74	0.74	0.66	0.65	0.46	0.45	0.36	0.32
bert-historical-rw	0.74	0.74	0.64	0.64	0.47	0.47	0.39	0.37
gbert-large	**0.77**	0.77	0.67	0.67	**0.51**	0.51	**0.40**	0.39
gelectra-large	**0.77**	0.77	**0.68**	0.67	**0.51**	0.51	0.39	0.36
bert-europeana-further-pretrained	0.65	0.57	0.54	0.43	0.29	0.23	0.19	0.12

Table 5: Evaluation results for the *speech corpus*. F1-scores are weighted F1. pol=*polarity*, t-p=*triple polarity*, m-c=*main class*, s-e=*sub-emotion*. Best result per classification is marked in bold for accuracies.

2019; Yang et al., 2019; Cao et al., 2020), however, for the most part, with larger training corpora and fewer classes as in our setting. We intend to improve the performance to satisfactory levels by hyperparameter-tuning and especially by exploring recommended ML-methods like over- and undersampling to deal with the class imbalances (Buda et al., 2018), which is one of the main problems of the *main class* and *sub-emotion* classification.

Among all transformer-based models, the *bert-europeana*-model further pretrained on dramatic texts yields the lowest accuracies. The performance becomes especially low for *main classes* and *sub-emotions* (see table 4). A reason might be that, while research argues that further pretraining with even low amount of texts can show improvements, the amount of text used in our setting (300 MB) is below the amounts reported in similar research (Kameswara Sarma et al., 2018; Gururangan et al., 2020; Rietzler et al., 2020). The usage of solely dramatic texts instead of varied forms of texts for training might also lead to problems in generalizing the specific language of the annotated material. Furthermore, a significant proportion of the selected dramatic texts is actually of the middle to the end of the 19th century and also of the beginning of the 20th century. Thus, the language might again deviate strongly from the time span of our plays (1745-1807). This might also be a reason why the historical transformer-based models in our evaluation show no relevant improvements. Investigating the training corpora for these models (Schweter, 2020; Brunner et al., 2020) shows that large proportions of the texts are actually of the 19th and 20th century. For our future studies, we plan to continue our exploration of domain-specific fine-tuning by acquiring larger amounts of general text material (and not only dramatic texts) focused on the time span of our interest, 1650-1815, to train models from scratch and evaluate if we can identify performance improvements. We intend to achieve satisfactory levels of accuracies to perform large-scale analysis of emotion distributions and progressions for our entire corpus of around 300 plays.

Acknowledgements

We want to thank the following student annotators for their contributions to this project: Carlina Eizenberger, Viola Hipler, Emma Ruß, Leon Sautter and Lisa Schattmann.

Funding

This research is part of the project "Emotions in Drama" (Emotionen im Drama), funded by the German Research Foundation (DFG) and part of the priority programme SPP 2207 Computational Literary Studies (CLS).

References

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional Sequencing and Development in Fairy Tales. In *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science, pages 668–674, Berlin, Heidelberg. Springer.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv:1903.10676 [cs]*. ArXiv: 1903.10676.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020. To bert or not to bert-comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In *SwissText/KONVENS*.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259. ArXiv: 1710.05381.

Lihong Cao, Sancheng Peng, Pengfei Yin, Yongmei Zhou, Aimin Yang, and Xinguang Li. 2020. A Survey of Emotion Analysis in Text Based on Deep Learning. In *2020 IEEE 8th International Conference on Smart City and Informatization (iSCI)*, pages 81–88, Guangzhou, China. IEEE.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. *arXiv:2010.10906 [cs]*. ArXiv: 2010.10906.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.

Diogo Cortiz. 2021. Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA. *arXiv:2104.02041 [cs]*. ArXiv: 2104.02041.

Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. 2020. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*, 9(3):483. ArXiv: 2006.03541.

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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jakob Fehle, Thomas Schmidt, and Christian Wolff. 2021. Lexicon-based sentiment analysis in german: Systematic evaluation of resources and preprocessing techniques. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, Düsseldorf, Germany.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. Conference Name: Digital Humanities 2019: "Complexities" (DH2019) Publisher: Zenodo.

Evelyn Gius, Jan Christoph Meister, Marco Petris, Malte Meister, Christian Bruck, Janina Jacke, Mareike Schuhmacher, Marie Flüh, and Jan Horstmann. 2020. CATMA.

Santiago González-Carvajal and Eduardo C. Garrido-Merchán. 2021. Comparing BERT against traditional machine learning text classification. *arXiv:2005.13012 [cs, stat]*. ArXiv: 2005.13012.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv:2004.10964 [cs]*. ArXiv: 2004.10964.

Prathusha Kameswara Sarma, Yingyu Liang, and Bill Sethares. 2018. Domain Adapted Word Embeddings for Improved Sentiment Classification. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 51–59, Melbourne. Association for Computational Linguistics.

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

Kai Labusch, Clemens Neudecker, and David Zellhofer. 2019. BERT for Named Entity Recognition in Contemporary and Historical German. page 9.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174. Publisher: [Wiley, International Biometric Society].

Saif Mohammad. 2011. From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained Models for Natural Language Processing: A Survey. *arXiv:2003.08271 [cs]*. ArXiv: 2003.08271.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31. ArXiv: 1606.07772.

Nils Reiter. 2020. *Anleitung zur Erstellung von Annotationsrichtlinien*, pages 193–202. De Gruyter.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Thomas Schmidt. 2019. Distant reading sentiments and emotions in historic german plays. In *Abstract Booklet, DH_Budapest_2019*, pages 57–60. Budapest, Hungary.

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. In Sandra Kübler and Heike Zinsmeister, editors, *Proceedings of the Workshop on Annotation in Digital Humanities 2018 (annDH 2018)*, pages 47–52. RWTH Aachen, Sofia, Bulgaria.

Thomas Schmidt, Manuel Burghardt, Katrin Dennerlein, and Christian Wolff. 2019a. Sentiment Annotation for Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts. In Thierry Declerck and John P. McCrae, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, pages 45–50. Leipzig, Germany.

Thomas Schmidt, Manuel Burghardt, and Christian Wolff. 2019b. Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, volume 2364 of *CEUR Workshop Proceedings*, pages 405–414, Copenhagen, Denmark. CEUR-WS.org.

Thomas Schmidt, Johanna Dangel, and Christian Wolff. 2021a. Senttext: A tool for lexicon-based sentiment analysis in digital humanities. In Thomas Schmidt and Christian Wolff, editors, *Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)*, volume 74, pages 156–172. Werner Hülsbusch, Glückstadt.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021b. Towards a Corpus of Historical German Plays with Emotion Annotations. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 9:1–9:11, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021c. Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. In Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, and Ulrike Wuttke, editors, *Fabrikation von Erkenntnis. Experimente in den Digital Humanities*.

Thomas Schmidt, Isabella Engl, David Halbhuber, and Christian Wolff. 2020a. Comparing live sentiment annotation of movies via arduino and a slider with textual annotation of subtitles. In *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*, pages 212–223.

Thomas Schmidt, Florian Kaindl, and Christian Wolff. 2020b. Distant reading of religious online communities: A case study for three religious forums on reddit. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*, pages 157–172, Riga, Latvia.

Thomas Schmidt, Brigitte Winterl, Milena Maul, Alina Schark, Andrea Vlad, and Christian Wolff. 2019c. Inter-rater agreement and usability: A comparative evaluation of annotation tools for sentiment annotation. In *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*, pages 121–133, Bonn. Gesellschaft für Informatik e.V.

Stefan Schweter. 2020. Europeana BERT and ELECTRA models.

Stefan Schweter and Johannes Baiter. 2019. Towards Robust Named Entity Recognition for Historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Boaz Shmueli and Lun-Wei Ku. 2019. SocialNLP EmotionX 2019 Challenge Overview: Predicting Emotions in Spoken Dialogues and Chats. *arXiv:1909.07734 [cs]*. ArXiv: 1909.07734.

Rachele Sprugnoli, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. 2015. Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31:762–772. Publisher: Oxford : Oxford University Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. ArXiv: 1910.03771.

Ian Wood, John McCrae, Vladimir Andryushechkin, and Paul Buitelaar. 2018a. A Comparison of Emotion Annotation Approaches for Text. *Information*, 9(5):117.

Ian Wood, John P. McCrae, Vladimir Andryushechkin, and Paul Buitelaar. 2018b. A Comparison Of Emotion Annotation Schemes And A New Annotated Data Set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kisu Yang, Dongyub Lee, Taesun Whang, Seolhwa Lee, and Heuiseok Lim. 2019. EmotionX-KU: BERT-Max based Contextual Emotion Classifier. *arXiv:1906.11565 [cs]*. ArXiv: 1906.11565.

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. Prediction of Happy Endings in German Novels. In *DMNLP@PKDD/ECML*.

Emily Öhman. 2020. Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pages 293–301. CEUR Workshop Proceedings.

A Appendix: Class Distributions for Corpus Manifestations

Emotion category	**absolute**	%
MC: emotions of affection	1,965	23
MC: emotions of joy	1,348	16
MC: emotions of fear	614	7
MC: emotions of rejection	2,153	25
MC: emotions of suffering	1,566	18
emotional movement	906	9
Overall	8,552	100

Table 6: Distributions of *main classes* for the *filtered corpus*. Percentages are rounded.

Emotion category	**absolute**	%
desire	28	0
love	1,032	14
friendship	185	2
admiration	468	6
joy	1,103	15
Schadenfreude	181	2
fear	390	5
despair	160	2
anger	1,002	13
hate, disgust	690	9
suffering	1,045	14
compassion	313	4
emotional movement	906	9
Overall	7,503	100

Table 7: Distributions of *sub-emotions* for the *filtered corpus*. *Polarity* distribution is 6,018 *negative* (60%) and 3,944 *positive* (40%). Percentages are rounded.

Emotion category	**absolute**	%
MC: emotions of affection	1,198	18
desire	27	0
love	602	9
friendship	126	2
admiration	441	7
MC: emotions of joy	1,088	16
joy	881	13
Schadenfreude	201	3
MC: emotions of fear	725	11
fear	391	6
despair	339	5
MC: emotions of rejection	1,538	23
anger	919	14
hate, disgust	660	10
MC: emotions of suffering	1,175	17
suffering	833	12
compassion	297	4
emotional movement	1,022	15
Overall	6,741	100

Table 8: Distributions of emotions for the *speech corpus*. *Polarity* distribution is 3,414 *negative* (60%) and 2,305 *positive* (40%). Percentages are rounded.