An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing

Thomas Schmidt
Media Informatics Group
Regensburg University
93040 Regensburg, Germany
thomas.schmidt@ur.de

Manuel Burghardt
Computational Humanities Group
Leipzig University
04109 Leipzig, Germany
burghardt@informatik.uni-leipzig.de

Abstract

We present results from a project on sentiment analysis of drama texts, more concretely the plays of Gotthold Ephraim Lessing. We conducted an annotation study to create a gold standard for a systematic evaluation. The gold standard consists of 200 speeches of Lessing's plays and was manually annotated with sentiment information by five annotators. We use the gold standard data to evaluate the performance of different German sentiment lexicons and processing configurations like lemmatization, the extension of lexicons with historical linguistic variants, and stop words elimination, to explore the influence of these parameters and to find best practices for our domain of application. The best performing configuration accomplishes an accuracy of 70%. We discuss the problems and challenges for sentiment analysis in this area and describe our next steps toward further research.

1 Introduction

As drama provides a number of structural features, such as speakers, acts or stage directions, it can be considered a literary genre that is particularly convenient and accessible for computational approaches. Accordingly, we find a number of quantitative approaches for the analysis of drama in general (cf. Fucks and Lauter, 1965; Solomon, 1971; Ilsemann; 2008), but also with a focus on the analysis of emotion (Mohammad, 2011) and sentiment (Nalisnick and Baird, 2013). More concretely, Mohammad (2011) uses the *NRC Emotion Lexicon* (Mohammad and Turney, 2010) to analyze the distribution and the progression of eight basic emotions in a selection of Shakespeare's plays. Nalisnick and Baird (2013) focus on speaker relations to analyze sentiment in Shakespeare's plays.

The goal of our study is to extend these existing approaches to computational sentiment analysis by taking into account historic, German plays. Further, we address some of the limitations of the current research on sentiment analysis in drama (Mohammad, 2011; Nalisnick and Baird, 2013), e.g. the ad hoc usage of sentiment lexicons without any pre-processing steps or other adjustments (Mohammad, 2011; Nalisnick and Baird, 2013). Our main contribution to the field of sentiment analysis for drama is the systematic evaluation of lexicon-based sentiment analysis techniques for the works of the German playwright Gotthold Ephraim Lessing. The evaluation takes into account a number of existing sentiment lexicons for contemporary German language (Võ et al., 2009; Clematide and Klenner, 2010; Mohammad and Turney, 2010; Remus et al., 2010; Waltinger, 2010) and various related NLP techniques, such as German lemmatizers, stop words lists and spelling variant dictionaries. The various combinations of existing lexicons and NLP tools are evaluated against a human annotated subsample, which serves as a gold standard.

2 Related Work: Sentiment Analysis in Literary Studies

As literary scholars have been interested in the emotions and feelings expressed in narrative texts for quite some time (cf. Winko, 2003; Mellmann, 2015), it is not surprising that computational sentiment analysis techniques have found their way into the realm of literary studies, for instance with Alm and Sproat (2005) and Alm et al. (2005), who examined the sentiment annotation of sentences in fairy tales. Other examples include Kakkonen and Kakkonen (2011), who used lexicon-based sentiment analysis to visualize and compare the emotions of gothic novels in a graph-based structure. Ashok et al. (2013) found a connection between the distribution of sentiment bearing words and the success of novels. Elsner (2012) included sentiment analysis to examine the plot structure in novels. In the same area, Jockers (2015) authored several blog posts about the use of sentiment analysis for the interpretation and visualization of plot arcs in novels. Reagan et al. (2016) extended Jockers work and use supervised as well as unsupervised learning to identify six core emotional arcs in fiction stories. Jannidis et al. (2016) used the results of lexicon-based sentiment analysis as features to detect "happy endings" in German novels of the 19th century. Heuser et al. (2016) used crowdsourcing, close reading and lexicon-based sentiment analysis to connect sentiments with locations of London in 19th century novels and visualize the information on maps of historic London. Kim et al. (2017) used lexical emotion features as part of a bigger feature set to successfully predict the genre of fiction books via machine learning. Buechel et al. (2016) identified the historical language of past centuries as a major challenge for sentiment analysis and therefore constructed a sentiment lexicon for the use case of 18th and 19th century German to analyze emotional trends and distributions in different genres.

3 Evaluation Design

3.1 Corpus

In order to investigate the practicability of lexicon-based sentiment analysis techniques for historical German drama, we gathered an experimental corpus of twelve plays by Gotthold Ephraim Lessing, which comprises overall 8,224 speeches. The plays were written between 1747 and 1779. Eight of the dramas are attributed to the genre of comedy while three are tragedies and one is referred to as dramatic poem. The most famous plays of the corpus are "*Nathan der Weise"* and "*Emilia Galotti"*. The average length of the speeches of the entire corpus is 24.15 words; the median is 13 words, which shows that the corpus consists of many rather short speeches. The longest speech consists of 775 words. Also note that the plays have very different lengths, with the shortest consisting of 183 and the longest of 1,331 speeches. All texts in our corpus are available in XML format and come with structural and speaker-related information for the drama text[1].

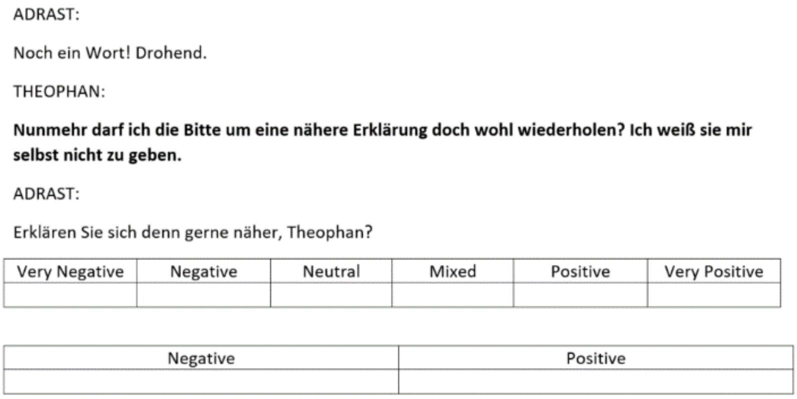
3.2 Gold Standard Creation

To be able to assess the quality of results from our evaluation study of lexicon-based approaches to sentiment analysis in Lessing's plays, we created a human annotated gold standard for 200 speeches. It is important to note that we were primarily interested in the overall sentiment of a self-contained character speech, as speeches are typically the smallest meaningful unit of analysis in quantitative approaches to the study of drama (cf. Ilsemann, 2008; Wilhelm, Burghardt and Wolff, 2013; Nalisnick and Baird, 2013). To create a representative sample of the 200 speeches, several characteristics of the corpus were taken into consideration: First, we only selected speeches longer than 18 words, which represents -25% of the average word length of speeches of the corpus, as we wanted to eliminate very short speeches that may contain no information at all. In related work, very short text snippets have been reported to be problematic for sentiment annotation, due to the lack of context and content (Alm and Sproat, 2005; Liu, 2016, p. 10). From the remaining speeches, we randomly selected speeches so that the proportion of speeches per drama in our gold standard represents the proportion per drama of the entire corpus, i.e. there are proportionally more speeches for longer dramas. We reviewed all speeches manually and replaced some speeches that consisted of French and Latin words, since those speeches might be problematic for our German speaking annotators. The final gold standard corpus had an average length of 50.68 words per speech and a median of 38 with the longest speech being 306 words long.

Five annotators, all native speakers in German, annotated the 200 speeches. During the annotation process, every speech was presented to the annotators with the preceding and the subsequent speech as contextual information. For the annotation scheme, we used two different approaches, which are oriented toward similar annotation studies (Bosco et al., 2014; Saif et al., 2014; Momtazi, 2012, Takala et al., 2014). First, annotators had to assign each speech to one of six categories: very negative, negative, neutral, mixed, positive, very positive. We refer to this annotation as *differentiated polarity*. In a second step, participants had to choose a *binary polarity*: negative or positive. This means, if annotators chose

[1] All electronic texts were gathered from the *TextGrid Repository* (https://textgridrep.org/repository.html).

neutral or *mixed* in the first step, they had to choose a binary polarity based on the overall tendency. With the first annotation, we wanted to gather some basic insights into sentiment distributions. However, several studies have shown that the agreement is very low for differentiated schemes like this (Momtazi, 2012; Takala et al., 2014); therefore, we also presented the binary annotation. Figure 1 illustrates the annotation scheme.



ADRAST:

Noch ein Wort! Drohend.

THEOPHAN:

Nunmehr darf ich die Bitte um eine nähere Erklärung doch wohl wiederholen? Ich weiß sie mir selbst nicht zu geben.

ADRAST:

Erklären Sie sich denn gerne näher, Theophan?

Very Negative	Negative	Neutral	Mixed	Positive	Very Positive

Negative	Positive

Figure 1. Example annotation

All five annotators had two weeks of time to conduct the entire annotation of 200 speeches, independently from each other. According to the annotators, the task took around five hours. An analysis of the differentiated annotations shows that the majority of annotations are *negative* or *very negative* (47%), while *positive* or *very positive* annotations are rather rare (16%). The results of the annotation also show that mixed (23%) and neutral (14%) annotations are a relevant annotation category as well. For the binary annotation, 67% of the annotations are negative and 33% are positive.

We analyzed the agreement of the annotations with *Krippendorff's α* (Krippendorff, 2011) and the average percentage of agreement of all annotator pairs (APA). Table 1 summarizes the results for the agreement of both annotation types.

	Krippendorff's α	APA
Differentiated polarity	0.22	40%
Binary polarity	0.47	77%

Table 1. Measures of agreement

Krippendorff's α and the average percentage of agreement of all annotator pairs point to a low agreement for the differentiated polarity. The degree of agreement is moderate for the binary polarity according to the interpretation of Landis and Koch (1971).

Since the degree of agreement is considerably higher for the binary polarity, we only regarded the binary polarity for the construction of our gold standard corpus. We selected the polarity chosen by the majority (>=3) of the annotators as the final value in our gold standard. This approach leads to 61 speeches being annotated as positive and 139 as negative. The entire gold standard corpus with all speeches, the final annotations and all other annotation data are publicly available[2].

3.3 Parameters of Evaluation

The results of automatic sentiment analysis approaches are influenced by a number of parameters. To find out which configuration of parameters yields the best results for historic plays in German language, we evaluated the following five variables:

[2] https://docs.google.com/spreadsheets/d/1f72hS2WDRBOrxzSY_tsM_igChG2bvxYTyMVZP6kOnuk/edit#gid=0

i) Sentiment lexicon

A sentiment lexicon is a list of words annotated with sentiment information. These words are also referred to as sentiment bearing words (*SBWs).* We identified five general purpose sentiment lexicons for German and evaluated their performance: *SentiWortschatz* (SentiWS, Remus et al., 2010), the *Berlin Affective Word List* (BAWL, Võ et al., 2009), *German Polarity Clues* (GPC; Waltinger, 2010), the German translation of the *NRC Emotion Lexicon* (NRC, Mohammad and Turney, 2010) and a sentiment lexicon by Clematide and Klenner (2010), further referred to as *CK*. Note, that all of the sentiment lexicons have different sizes and were created in different ways. They also differ in their overall composition: Some have simple binary polarity annotations, i.e. a word is either positive or negative. We refer to this kind of annotation as *dichotomous polarity*. Others have additional polarity strengths, which are values on a continuous scale e.g. ranging from -1 (very negative) to +1 (very positive) (SentiWS, CK, BAWL). Most of the lexicons consist of the base forms of words (lemmas), but some are manually appended with inflections of the words (SentiWS, GPC). Besides the lexicons, we also created and evaluated a combination of all five lexicons. To do this, we simplified the basic idea of sentiment lexicon combination by Emerson and Declerk (2014), i.e. we merged all words of all lexicons. If words were annotated ambiguously, we selected the polarity annotation that occurred in the majority of lexicons. For this process, we only regarded the dichotomous polarity of the lexicons.

ii) Extension with linguistic variants

The development of the aforementioned lexicons is based on modern online lexicons (Võ et al., 2009), corpora of product reviews (Remus et al., 2010), news articles (Clematide and Klenner, 2010) and the usage of crowdsourcing (Mohammad and Turney, 2010). Therefore, the lexicons were rather created to be used for contemporary language than for poetic German language of the 18th century. Some early studies in this area already identified the problem of historical language for contemporary sentiment lexicons (Alm and Sproat, 2005; Sprugnoli et al., 2016; Buechel et al., 2016). To examine this problem, we used a tool of the *Deutsches Textarchiv* (DTA) that produces historical linguistic variants of German words, e.g. different orthographical variants a word had throughout history (Jurish, 2012). The tool also provides historical inflected forms of the words. We used this tool to extend the sentiment lexicons as we gathered all historical linguistic variants for each word of every lexicon and added those words to the lexicon with the same polarity annotation of the base. This procedure increased the size of the lexicons to a large degree, since for every orthographic variant all inflections were added (example: size of BAWL before extension: 2,842; after extension: 75,436). However, one of the dangers of this approach is the possible addition of words that are not really sentiment bearing words, which may skew the polarity calculation. Further, the DTA tool has not yet been evaluated for this specific use case, so the quality of the produced variants is unclear. Hence, we evaluate the performance of the lexicons in their basic form (*noExtension*) as well as with the extension (*dtaExtended*).

iii) Stop words lists

We also analyzed the influence of stop words and frequently occurring words of the corpus on the sentiment calculation. Saif et al. (2014) showed that the elimination of these words can have a positive influence on performance of sentiment analysis in the machine learning context. Stop words and frequent words might skew sentiment calculation of lexicon-based methods as well, since some of the lexicons actually contain stop words. There are also some highly frequent words in our corpus that are listed in many of the lexicons as sentiment bearing words, but are actually overused because of the particular language style of the 18th century. We use different types of stop words lists to explore the influence of those types of words:

- a basic German stop word list of 462 words (upper and lower case*; standardList*),
- the same list extended by the remaining 100 most frequent words of the entire Lessing corpus (*extendedList*),
- and the same list manually filtered by words that are indeed very frequent, but are still sentiment bearing (e.g. Liebe/love; *filteredExtendedList*).

Besides, we also evaluate the condition to use no stop words list at all (*noStopWords*).

iv) Lemmatizers

We evaluate lemmatization by using and comparing two lemmatizers for German: the *treetagger* by Schmidt (1995) and the *pattern lemmatizer* by De Smedt and Daelemans (2012). Many of the lexicons only include base forms of SBWs, so lemmatization is a necessary step for those lexicons to identify inflections. However, due to the general problems of automatic lemmatization in German (Eger et al., 2016) and the special challenges historical and poetic language pose to automatic lemmatizers, mistakes and problems might occur that distort the detection of SBWs. Besides the general comparison between the two lemmatizers and no lemmatization at all, we also compared the automatic lemmatization with the manually added inflections some lexicons contain and the extension with inflections by the tool of the DTA.

v) Case-sensitivity

Related studies with lexicon-based methods typically lowercase all words for reasons of processing and normalization (Klinger et al., 2016). We wanted to explore if case affects the evaluation results (*caseSensitive* vs. *caseInSensitive*), as several words in German have a change in meaning depending on case, especially with regard to their sentiment.

3.4 Sentiment Calculation

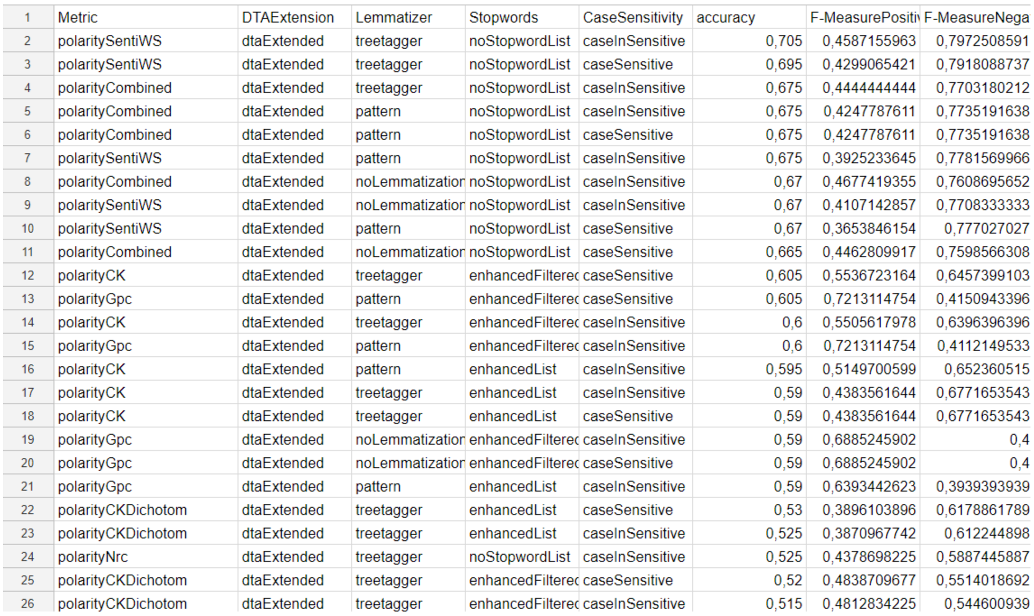
To calculate the sentiment of a speech we employed a simple term counting method, often used with lexicon-based methods (Kennedy and Inkpen, 2006). The number of positive words according to the used configuration of parameters is subtracted by the number of negative words to get a general polarity score. If this score is negative, the speech is regarded as negative, otherwise as positive. If a sentiment lexicon contains polarity strengths for the SBWs, we additionally used these values in a similar way to calculate a sentiment score. Therefore, for these lexicons we calculated and compared two scores: one for dichotomous polarity and one for polarity strengths.

4 Results

In this chapter we present results from our evaluation of all possible combinations of the previously described parameters in comparison to the human annotated gold standard data. We used well-established metrics for sentiment analysis evaluation (Gonçalves et al., 2013). Our main metric is *accuracy*, which is the proportion of the correctly predicted speeches of all speeches. To get a more holistic view of the results, we also looked at *recall*, *precision* and *F-measures*. We furthermore analyzed the metrics for positive and negative speeches separately. Since our gold standard has an overrepresentation of negative speeches, misbehaviors of configurations like a general prediction of negative speeches would otherwise go undetected. We use the random baseline, the majority baseline and agreement measures as benchmarks. Because of the unequal distribution, the random baseline is set to 0.525 and the majority baseline is set to 0.695 for the accuracy. Mozetic et al. (2016) propose the average percentage of agreement of all annotator pairs (APA, 77%) as baseline for sentiment analysis evaluations. We also take this baseline into account when assessing the performance.

As we analyzed more than 400 different configurations, we are not able present all evaluation results in detail in this paper. However, an evaluation table of the best configurations ordered by accuracy is available online[3]. Note that we removed configurations from the table that tend to predict almost all speeches as negative and therefore accomplish accuracies close to the majority baseline, but are actually flawed. Figure 2 shows a snippet of the top part of the table.

[3] https://docs.google.com/spreadsheets/d/1yOv0U99SDI0dFUkctJcGmTXHSxcRByRJihnqRhEIkNY/edit#gid=0



1	Metric	DTAExtension	Lemmatizer	Stopwords	CaseSensitivity	accuracy	F-MeasurePositi	F-MeasureNega
2	polaritySentiWS	dtaExtended	treetagger	noStopwordList	caseInSensitive	0,705	0,4587155963	0,7972508591
3	polaritySentiWS	dtaExtended	treetagger	noStopwordList	caseSensitive	0,695	0,4299065421	0,7918088737
4	polarityCombined	dtaExtended	treetagger	noStopwordList	caseInSensitive	0,675	0,4444444444	0,7703180212
5	polarityCombined	dtaExtended	pattern	noStopwordList	caseInSensitive	0,675	0,4247787611	0,7735191638
6	polarityCombined	dtaExtended	pattern	noStopwordList	caseSensitive	0,675	0,4247787611	0,7735191638
7	polaritySentiWS	dtaExtended	pattern	noStopwordList	caseInSensitive	0,675	0,3925233645	0,7781569966
8	polarityCombined	dtaExtended	noLemmatizatior	noStopwordList	caseInSensitive	0,67	0,4677419355	0,7608695652
9	polaritySentiWS	dtaExtended	noLemmatizatior	noStopwordList	caseInSensitive	0,67	0,4107142857	0,7708333333
10	polaritySentiWS	dtaExtended	pattern	noStopwordList	caseSensitive	0,67	0,3653846154	0,777027027
11	polarityCombined	dtaExtended	noLemmatizatior	noStopwordList	caseSensitive	0,665	0,4462809917	0,7598566308
12	polarityCK	dtaExtended	treetagger	enhancedFilterec	caseSensitive	0,605	0,5536723164	0,6457399103
13	polarityGpc	dtaExtended	pattern	enhancedFilterec	caseSensitive	0,605	0,7213114754	0,4150943396
14	polarityCK	dtaExtended	treetagger	enhancedFilterec	caseInSensitive	0,6	0,5505617978	0,6396396396
15	polarityGpc	dtaExtended	pattern	enhancedFilterec	caseInSensitive	0,6	0,7213114754	0,4112149533
16	polarityCK	dtaExtended	pattern	enhancedList	caseInSensitive	0,595	0,5149700599	0,652360515
17	polarityCK	dtaExtended	treetagger	enhancedList	caseInSensitive	0,59	0,4383561644	0,6771653543
18	polarityCK	dtaExtended	treetagger	enhancedList	caseSensitive	0,59	0,4383561644	0,6771653543
19	polarityGpc	dtaExtended	noLemmatizatior	enhancedFilterec	caseInSensitive	0,59	0,6885245902	0,4
20	polarityGpc	dtaExtended	noLemmatizatior	enhancedFilterec	caseSensitive	0,59	0,6885245902	0,4
21	polarityGpc	dtaExtended	pattern	enhancedList	caseInSensitive	0,59	0,6393442623	0,3939393939
22	polarityCKDichotom	dtaExtended	treetagger	enhancedList	caseSensitive	0,53	0,3896103896	0,6178861789
23	polarityCKDichotom	dtaExtended	treetagger	enhancedList	caseInSensitive	0,525	0,3870967742	0,612244898
24	polarityNrc	dtaExtended	treetagger	noStopwordList	caseInSensitive	0,525	0,4378698225	0,5887445887
25	polarityCKDichotom	dtaExtended	treetagger	enhancedFilterec	caseSensitive	0,52	0,4838709677	0,5514018692
26	polarityCKDichotom	dtaExtended	treetagger	enhancedFilterec	caseInSensitive	0,515	0,4812834225	0,544600939

Figure 2. Table snippet of the results of the evaluation of all configurations

Table 2 reduces the results to the best configurations of every single sentiment lexicon and shows the corresponding accuracies as well as F-measures for both polarity classes (positive/negative).

lexicon	extension	lemmatization	stop words	case	accuracy	F-Positive	F-Negative
SentiWS (Polarity Strengths)	dtaExtended	treetagger	noStopWords	caseInSensitive	0.7	0.46	0.79
Combined Lexicon	dtaExtended	treetagger	noStopWords	caseInSensitive	0.68	0.44	0.77
CK (Polarity Strengths)	dtaExtended	treetagger	enhancedFilteredList	caseSensitive	0.6	0.55	0.65
GPC	dtaExtended	pattern	enhancedFilteredList	caseSensitive	0.6	0.72	0.42
NRC	dtaExtended	treetagger	noStopWords	caseInSensitive	0.53	0.44	0.59
BAWL	dtaExtended	treetagger	noStopWords	caseInSensitive	0.49	0.50	0.46

Table 2. Best configurations per sentiment lexicon

In the following we summarize the main findings of our evaluation study:

- The overall best performance is delivered by the SentiWS lexicon and the combined lexicon if the remaining parameters are the same
- When a lexicon has polarity strengths, calculation with those always outperform calculations with the dichotomous polarity of the lexicon. Apart from BAWL, the general rule is that lexicons with polarity strengths (SentiWS, CK) in general outperform all other lexicons with only dichotomous polarities (GPC, NRC)
- The extension with historical linguistic variants consistently yields the strongest performance boost for all lexicons. The extension with historical inflections is better than just automatic lemmatization.

- Stop words lists have differentiated influences. Some lexicons (e.g. GPC) tend to excessive prediction of one polarity class, because of stop words and frequent words. However, this does not always lead to worse accuracies but in-depth word analysis shows incorrect sentiment assignments to words. We therefore recommend the usage of stop word lists when lexicons contain stop words or when stop words are generated by means of additional NLP processes. The best rated lexicons (e.g. SentiWS) are not influenced by stop words at all.
- Both lemmatizers perform almost equally good. A detailed analysis of the results shows that both lemmatizers have problems with the historical language of the speeches. However, for lexicons that consist only of base forms, lemmatization leads to a better overall performance. The results for lexicons with manually added inflections show that those inflections work better than the automatic lemmatization. For many lexicons, a combination of both yields better results.
- Case-sensitivity does not have an effect on the overall quality of sentiment evaluations in our test corpus.

The best overall performance is accomplished with the polarity strengths of the SentiWS lexicon extended with historical linguistic variants, lemmatization via *treetagger*, no stop words lists and ignoring case-sensitivity. The accuracy for this performance is 0.705 with 141 speeches correctly predicted. This result is over the benchmark of the random and majority baseline but below the average percentage of agreement of the annotator pairs (77%).

5 Discussion and Outlook

We made several contributions to the research area of sentiment analysis in historical drama texts, one being the creation of an annotated corpus of drama speeches. However, there are some limitations concerning the annotation study: We identified low to mediocre levels of agreement among the annotators for the polarity annotation. The low level of agreement for annotation schemes with multiple categories is also found in several other research areas (Momtazi, 2012; Takala et al., 2014). The mediocre levels of agreement for the binary annotation are in line with similar research in the field of literary texts (Alm and Sproat, 2005; Alm et al., 2005) and texts with historical language (Sprugnoli et al., 2016). However, compared to other types of text, the agreement for the binary polarity is rather low (e.g. Thet et al., 2010; Prabowo and Thelwall, 2009). Sentiment annotation of literary texts seems to be a rather subjective and challenging task. Our annotators also reported difficulties due to the lack of context and general problems in understanding the poetic and historical language of Lessing. Note that many of the challenges very likely occurred because the annotators were non-experts concerning the drama texts. Some feedback of the annotators also points to the possibility that the used annotation schemes were not sufficient or representative for the application area of sentiment in historical plays. Another limitation is the small size of the corpus: 200 speeches amount to 2% of the speeches of our original Lessing corpus. While such small sample sizes are not uncommon for sentiment annotation of literary texts (Alm and Sproat, 2005), they certainly lessen the significance of the results. To address some of the mentioned limitations, we are planning to conduct larger annotation studies with trained experts in the field of Lessing, more speeches and a more sophisticated annotation scheme.

Our major contribution is the systematic evaluation of different configurations of lexicon-based sentiment analysis techniques. Many of our findings are important not only for sentiment analysis of German drama texts, but for sentiment analysis of corpora with historical and poetic language in general. We identified SentiWS (Remus et al., 2010) as the best performing lexicon for our corpus. The accuracy of the combined lexicon is overall slightly lower than the top rated SentiWS-configuration. The reason for this might be the extension of SentiWS by many problematic and distorting sentiment bearing words that can be regarded as noise. Furthermore, the transfer of some problems of other lexicons, like the missing of manually added inflections, may be responsible for the decreased performance of the combined lexicon as compared to SentiWS alone.

We highly recommend the usage of sentiment lexicons with polarity strengths, since they consistently outperform dichotomous polarity calculations. This proves that for calculation purposes, sentiment bearing words are better represented on a continuous scale. The calculation with polarity strengths seems to better represent human sentiment annotation than the usage of dichotomous values (+1 / -1), as many sentiment bearing words indeed have different intensities that are perceived differently by human annotators and therefore should also be weighted differently for the automatic sentiment calculation.

The noticeable and consistent performance boost from the extension of lexicons by historical linguistic variants highlights the linguistic differences between the contemporary language of the lexicons and the historical language of the 18th century. The creation of sentiment lexicons, especially for the historical language of past centuries, is a beneficial next step and possibilities for historical German are already examined (Buechel et al., 2016). Considering stop words, we highly recommend checking sentiment lexicons for highly frequent words of the corpus, especially for historical and poetic language. The meaning concerning the sentiment of words can differ throughout history (Hamilton et al., 2016) and is also dependent on the linguistic style of a specific author. Therefore, stop words and other highly frequent words might have different sentiment connotations in contemporary and historic German language. It also shows that historic language poses challenges for automatic lemmatization, as it is not as effective as the extension by historical inflections.

Overall, we were able to achieve acceptable levels of accuracy with our best performing configuration (70%), considering the basic methods used, the linguistic challenges the corpus poses and the mediocre levels of agreement of the annotators. Furthermore, we did not consider sentiment classes like *neutral* or *mixed*, although the annotation study showed that many speeches of the corpus are actually not strictly *positive* or *negative*. However, accuracy results in other application areas of sentiment analysis like product reviews or social media are generally higher, e.g. around 90% (Vinodhini and Chandrasekaran, 2012). Besides the historical and poetic language difficulties, common problems of lexicon-based methods like the handling of irony and negations are certainly additional reasons for the mediocre accuracies. Based on our results, we consider the usage of general purpose lexicons alone as not sufficient to achieve acceptable accuracy scores. Using the results of the planned large-scale annotation studies, we will try to create corpora for evaluation and the development of more sophisticated methods of sentiment analysis, such as machine learning and hybrid techniques in order to improve accuracy and integrate other polarity classes as well.

We are also aware that more complex emotional categories like anger, trust or surprise are also of interest for sentiment analysis in literary texts (Alm and Sproat, 2005; Mohammad, 2011). While at the moment resources and best practices to include emotional categories in sentiment analysis are rare (Mohammad and Turney, 2010), we are expecting substantial progress from an ongoing shared task on implicit emotion recognition to gather more insights into this area[4].

Another limitation of the presented results is that we only regarded speeches for sentiment analysis. However to further explore the possibilities and use cases for sentiment analysis on drama texts, we developed a web tool[5] for the exploration of sentiment in Lessing's plays. The tool visualizes the results of the best performing configuration of our evaluation study. Literary scholars can explore sentiment polarity distributions and progressions on several levels, i.e. for the whole play, for single acts, scenes and speeches, but also for individual speakers or for the relationship between two speakers. Further, we also integrated the results of sentiment calculation with the NRC Emotion Lexicon (Mohammad and Turney, 2010), so besides polarity, more complex emotion categories like anger and surprise can be explored as well.

As an example, Figure 3 illustrates the polarity progression for Lessing's "Emilia Galotti" throughout the five acts of the play. On the x-axis, every act is represented by a bar. On the y-axis the polarity of the entire act is represented as an absolute value. The tool also enables the analysis of normalized values, e.g. by the length of the text unit. The tool shows that based on our polarity calculation the play starts with a rather positive polarity in the first act, but becomes more and more negative as the play progresses.

[4] More information about this shared task: http://implicitemotions.wassa2018.com/

[5] http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/FrontEnd/sa_selection.html

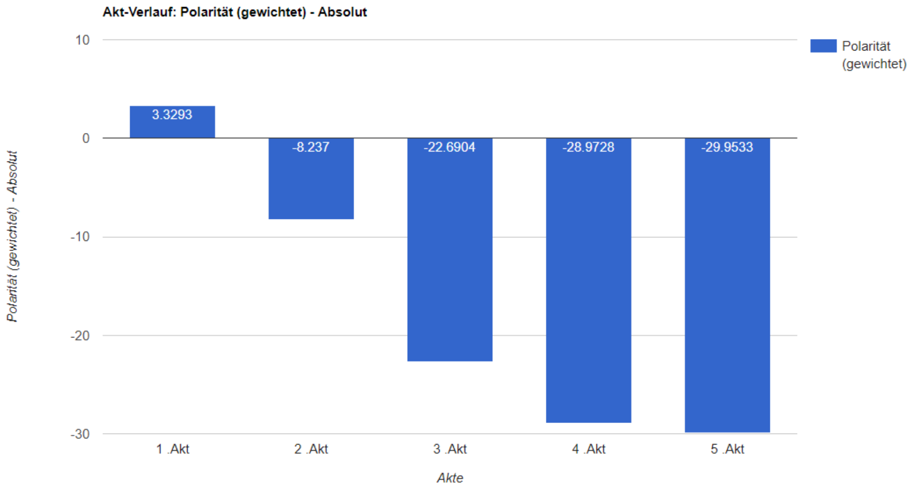


Figure 3. Polarity progression for Emilia Galotti per Act

This tool only represents a first prototype. Meanwhile, we are working close with literary scholars to gather more insights into needs and requirements for the literary analysis of emotion and sentiment in drama texts. By extending our corpus to other authors and eras, we also plan to explore sentiment analysis on drama texts beyond Lessing's plays.

Reference

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 579-586). Association for Computational Linguistics

Alm, C. O. & Sproat, R. (2005). Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 668-674). Springer Berlin Heidelberg.

Ashok, V. G., Feng, S., & Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1753-1764).

Bosco, C., Allisio, L., Mussa, V., Patti, V., Ruffo, G., Sanguinetti, M., & Sulis, E. (2014). Detecting happiness in Italian tweets: Towards an evaluation dataset for sentiment analysis in Felicitta. In *Proc. of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Opena Data, ESSSLOD* (pp. 56-63).

Buechel, S., Hellrich, J., & Hahn, U. (2016). Feelings from the Past – Adapting Affective Lexicons for Historical Emotion Analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (pp. 54-61).

Clematide, S. & Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. *In Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 7-13).

De Smedt, T. & Daelemans, W. (2012). *Pattern for Python. Journal of Machine Learning Research*, 13, 2031–2035.

Eger, S., Gleim, R., & Mehler, A. (2016). Lemmatization and Morphological Tagging in German and Latin: A Comparison and a Survey of the State-of-the-art. In *LREC*.

Elsner, M. (2012). Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 634-644). Association for Computational Linguistics.

Emerson, G. & Declerck, T. (2014). SentiMerge: Combining sentiment lexicons in a Bayesian framework. In *Proceedings of the Workshop on Lexical and Grammatical Re-sources for Language Processing* (pp. 30-38).

Fucks, W. & Lauter, J. (1965). Mathematische Analyse des literarischen Stils. In Kreuzer, H. & Gunzenhäuser, F. (Hrsg.), *Mathematik und Dichtung*, (pp. 107-122). München: Nymphenburger Verlagshandlung

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27-38). ACM.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, p. 595). NIH Public Access.

Heuser, R., Moretti, F., & Steiner, E. (2016). *The emotions of London*. Retrieved from https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf

Ilsemann, H. (2008). More statistical observations on speech lengths in Shakespeare's plays. *Literary and Linguistic Computing*, 23(4), 397-407.

Jannidis, F., Reger, I., Zehe, A., Becker, M., Hettinger, L. & Hotho, A. (2016*). Analyzing Features for the Detection of Happy Endings in German Novels.* arXiv preprint arXiv:1611.09028.

Jockers, M. L. (2015). R*evealing sentiment and plot arcs with the syuzhet package*. Retrieved from http://www.matthewjockers.net/2015/02/02/syuzhet/

Jurish, B. (2012*). Finite-state Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam (defended 2011). URN urn:nbn:de:kobv:517-opus-55789.

Kakkonen, T. & Kakkonen, G. G. (2011). SentiProfiler: creating comparable visual profiles of sentimental content in texts. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage* (pp. 62-69).

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2), 110-125.

Kim, E., Padó, S., & Klinger, R. (2017). Prototypical Emotion Developments in Literary Genres. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 17–26).

Klinger, R., Suliya, S. S., & Reiter, N. (2016). Automatic Emotion Detection for Quantitative Literary Studies. In *Digital Humanities Book of Abstracts 2016.*

Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. Retrieved from http://repository.upenn.edu/asc_papers/43

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Liu, B. (2016). *Sentiment Analysis. Mining Opinions, Sentiments and Emotions.* New York: Cambridge University Press.

Mellmann, K. (2015). Literaturwissenschaftliche Emotionsforschung. In: Rüdiger Zymner (Hg.): *Handbuch Literarische Rhetorik*. Berlin/Boston, 173-192.

Mohammad, S. (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* (pp. 105-114). Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34). Association for Computational Linguistics.

Momtazi, S. (2012). Fine-grained German Sentiment Analysis on Social Media. In *LREC* (pp. 1215-1220).

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PloS one*, 11(5), e0155036.

Nalisnick, E. T. & Baird, H. S. (2013). Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 479–483).

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), 143-157.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1), 31.

Remus, R., Quasthoff, U. & Heyer, G. (2010). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In *LREC* (pp. 1168-1171).

Saif, H., Fernandez, M., He, Y., Alani, H. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In: *Proc. 9th Language Resources and Evaluation Conference (LREC)* (pp. 810-817).

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop.*

Solomon, M. (1971). Ein mathematisch-linguistisches Dramenmodell. *Zeitschrift für Literaturwissenschaft und Linguistik,* 1(1), 139-152.

Sprugnoli, R., Tonelli, S., Marchetti, A., & Moretti, G. (2016). Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4), 762-772.

Takala, P., Malo, P., Sinha, A., & Ahlgren, O. (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. In *LREC* (Vol. 2014, pp. 2152-2157).

Thet, T. T., Na, J. C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6), 823-848.

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6), 282-292.

Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior research methods*, 41(2), 534-538.

Waltinger, U. (2010). Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10)*

Wilhelm, T., Burghardt, M., & Wolff, C. (2013). “To See or Not to See” - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In R. Franken-Wendelstorf, E. Lindinger, & J. Sieck (Eds.), *Kultur und Informatik: Visual Worlds & Interactive Spaces* (pp. 175–185). Glückstadt: Verlag Werner Hülsbusch.

Winko, S. (2003). Über Regeln emotionaler Bedeutung in und von literarischen Texten. In: Fotis Jannidis & Gerhard Lauer & Matias Martinez & SW (eds.): *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte.* Berlin, New York: de Gruyter, 329-348.