Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior

Thomas Schmidt[1], Manuel Burghardt[2], Katrin Dennerlein[3]
[1] Media Informatics Group, Regensburg University, 93040 Regensburg, Germany
[2] Computational Humanities Group, Leipzig University, 04109 Leipzig, Germany
[3] Department of German Philology, Würzburg University, 97074 Würzburg, Germany

thomas.schmidt@ur.de, burghardt@informatik.uni-leipzig.de, katrin.dennerlein@uni-wuerzburg.de

Abstract

We present results of a sentiment annotation study in the context of historical German plays. Our annotation corpus consists of 200 representative speeches from the German playwright Gotthold Ephraim Lessing. Six annotators, five non-experts and one expert in the domain, annotated the speeches according to different sentiment annotation schemes. They had to annotate the differentiated polarity (very negative, negative, neutral, mixed, positive, very positive), the binary polarity (positive/negative) and the occurrence of eight basic emotions. After the annotation, the participants completed a questionnaire about their experience of the annotation process; additional feedback was gathered in a closing interview. Analysis of the annotations shows that the agreement among annotators ranges from low to mediocre. The non-expert annotators perceive the task as very challenging and report different problems in understanding the language and the context. Although fewer problems occur for the expert annotator, we cannot find any differences in the agreement levels among non-experts and between the expert and the non-experts. At the end of the paper, we discuss the implications of this study and future research plans for this area.

Keywords: sentiment analysis, sentiment annotation, drama

1. Introduction

The analysis of emotions, affects, moods, feelings and sentiments in literary texts and their effect on the reader has a long hermeneutical tradition in literary studies (Winko, 2003; Meyer-Sickendiek, 2005; Mellmann, 2015). Lately, this area of study has been enhanced by computational sentiment analysis techniques, which are used to automatically predict sentiments and emotions in written texts (cf. Alm et al., 2005; Volkova et al., 2010; Jannidis et al., 2016; Kakkonen & Kakkonen, 2011; Kao & Jurafsky, 2012; Mohammad, 2011; Nalisnick & Baird, 2013; Schmidt et al., 2018). Sentiment analysis has become one of the most active areas of research in computational linguistics in recent years (Vinodhini & Chandrasekran, 2012) and is typically used for the analysis of online reviews and social media (Liu, 2016). However, a major problem for the application of sentiment analysis methods for literary texts is the lack of human-annotated training data. Such data is an important prerequisite for the evaluation of dictionary-based approaches (lists of words annotated with sentiment information), which are among the most popular methods for the sentiment analysis of literary texts (Mohammad, 2011; Nalisnick & Baird, 2013; Schmidt et al., 2018). Manually curated training data is even more important for unsupervised machine learning approaches, which have been proven to be very successful in the context of other areas of sentiment analysis (Pang et al., 2002).

Not only is there a lack of available training data; we currently also lack research concerning difficulties and problems in the transfer of standard methods for sentiment annotation (mostly used in online reviews and social media) to the field of narrative texts. For the area of fairy tales, Alm and Sproat (2005) conducted annotation studies and reported several problems, such as low agreement among annotators, strong imbalances concerning the distribution of sentiments and misinterpretations of the sentiment annotation scheme. Another question that arises, is the level of expertise necessary to correctly annotate sentiment: In the context of historical political texts, Sprugnoli et al. (2016) have found strong differences in annotations among experts, among participants of a crowdsourcing project and between the experts and the crowd. Furthermore, the special needs in sentiment annotation as well as requirements concerning the analysis of sentiments and emotions in narrative and poetic texts have yet to be explored. Sprugnoli et al. (2016) were able to identify special interests of professional historians for sentiment analysis and annotation (e.g. the sentiment of specific topics rather than text units) by including them in the annotation process.

As a prerequisite for a large-scale automatic annotation project, we are currently exploring sentiment annotation for historic (18th century) plays by G. E. Lessing, to examine the aforementioned questions and challenges concerning sentiment annotation of German, literary texts. In this article, we present preliminary annotation results of our first experiments with five non-expert annotators and one expert annotator with a corpus of 200 speeches. In addition, we also used a questionnaire and conducted interviews with the annotators to gather more insights concerning the annotation behavior as well as problems with the annotation scheme and the overall process. With regard to our overall project, we want to derive specific requirements for sentiment annotation in literary studies and examine which level of expertise is necessary for this specific context. Thus, we want to aid the development of annotation schemes and annotation tools for this area and further support the planning of future annotation studies.

2. Methods

The corpus of the overall project consists of twelve plays and altogether 8,224 speeches with an average length of 24 words per speech. For our annotation study, we randomly selected a sample of 200 speeches. Five non-expert annotators (four female and one male) participated in the study. They were all fluent in German but otherwise no experts concerning the plays of Lessing. One expert annotator (female) with a PhD in German literary studies

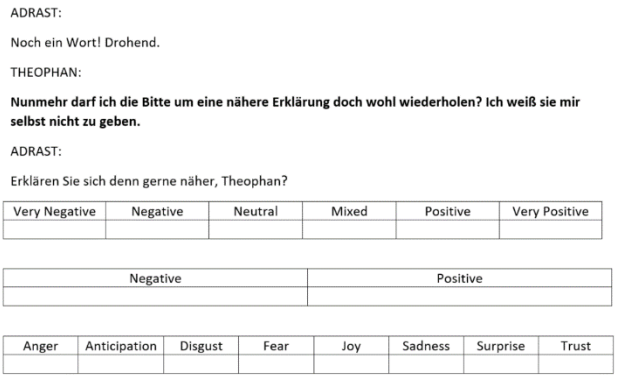
Please cite as:
Schmidt, T., Burghardt, M. & Dennerlein, K. (2018). Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior. In: Sandra Kübler, Heike Zinsmeister (eds.), *Proceedings of the Workshop on Annotation in Digital Humanities (annDH 2018)* (pp. 47-52). Sofia, Bulgaria.

and with research experience especially about Lessing also participated in the study. With this sample, we were able to gather a total of 1,200 annotations.

Since very short speeches may not contain any sentiment bearing words at all and generally pose challenges for the annotators due to a lack of context, we only selected speeches with a minimum length of 19 words, which equals about -25% of the average speech length. In the final annotation corpus, speeches had an average length of 50 words. Furthermore, we selected the speeches to reflect the distribution of speeches for different plays in our corpus, i.e. plays with overall more speeches are also represented with more speeches in our test corpus. We excluded speeches from our test corpus when we assumed language issues for the annotators, for instance speeches containing French or Latin words, which may be problematic for the German speaking annotators. Note that 200 speeches represent approx. 2% of our entire corpus. Although this might be considered a rather small sample size, this is not uncommon for the domain of historical and poetic texts (cf. Alm & Sproat, 2005; Sprugnoli, 2016), as annotations of this type are typically a laborious task.

The annotators were asked to use a multi-part annotation scheme based on various existing schemes for sentiment analysis. Most related studies use a categorical annotation scheme, differentiating only positive, negative, neutral / objective, mixed and unknown (Bosco et al., 2014; Refaee & Rieser, 2014; Saif et al., 2013). Other studies refer to ordinal or continuous ratings, ranging from positive to negative (Takala et al., 2014; Momtazi, 2012). Wiebe et al. (2005) developed a more complex scheme consisting of polarity categories and intensities for these categories. However, related work shows that oftentimes initially more sophisticated schemes are later simplified to a binary variant (positive/negative), since more complicated schemes cause lower agreement between human annotators (Momtazi, 2012; Takala et al., 2014). This reduction can also be observed in literary studies: Alm and Sproat (2005) at first used a complex annotation scheme with different emotional categories but then reduced it to a binary polarity of "emotion present" and "emotion not present" (Alm et al., 2005). Sprugnoli et al. (2016) chose a basic scheme of positive, negative, neutral and unknown.

In our study, we wanted to investigate whether this observation is also true for historic German plays, asking the annotators to use both, a fairly simple scheme and a more complex annotation scheme. The annotators were presented each of the 200 speeches together with the predecessor and successor speech, to provide the necessary context for interpretation. First, annotators were asked to assign one of six categories (very negative, negative, neutral, mixed, positive and very positive) to each speech. We will refer to this annotation as *differentiated polarity annotation.* Next, they had to assign a binary annotation (pos/neg). Finally, participants were able to annotate the presence of one or more emotion categories from a set of eight basic emotions (anger, fear, surprise, trust, anticipation, joy, disgust, sadness). Figure 1 illustrates the annotation process:


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

Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust

Figure 1: Example annotation task

For the differentiated polarity and the binary polarity, every annotator was asked to choose the most adequate sentiment category. For the emotion category, the instruction was to mark any emotions that are present in a speech. Every annotator was personally introduced to the annotation process, which was also explained with practical examples. At the end of the overall annotation task, participants were asked to complete a questionnaire about different facets of the annotation process. In the first part of the questionnaire, participants rated their overall impression of the annotation tasks on a 7-point Likert scale (do not agree at all/fully agree):

- The annotation of the speeches was difficult. (overall-difficulty)
- The annotation of the speeches concerning the polarity was difficult. (polarity-difficulty)
- The annotation of the speeches concerning the emotion categories was difficult. (emotion-difficulty)
- I was very confident with my assignments. (overall-certainty)
- I was very confident with my assignments concerning the polarities of the speeches. (polarity-certainty)
- I was very confident with my assignments concerning the emotion categories of the speeches. (emotion-certainty)

In addition, participants were asked to report how much time they needed to perform the annotation of all 200 speeches. Annotators were also asked to report about the most important problems and difficulties in a free response field. Finally, we conducted a short closing interview with all participants after the complete annotation task, discussing their overall experience with the annotation process.

3. Results

As differences and similarities between the annotations of non-experts and the domain expert are of special interest for us with regard to the design of future annotation studies, we will examine these data sets separately. Firstly, we report the results concerning distributions for the differentiated polarity annotation among non-experts in Table 1 (in total 1000 annotations).

Polarity	Number of annotations (Percentage)
Very Negative	99 (10%)
Negative	371 (37%)
Neutral	137 (14%)
Mixed	227 (23%)
Positive	133 (13%)
Very Positive	33 (3%)

Table 1: Frequency distribution of polarity annotations among non-experts.

We observed that for the vast majority of annotations our participants chose negative annotations. They represent almost 50% of all annotations, while the share of positive and very positive annotations is significantly lower (16%). The results also show that the groups "mixed" and "neutral" are relevant and important annotation groups, since they appear almost as often or even more frequently than positive annotations. As for the binary polarity, we found that 665 annotations (67%) were negative and 335 (33%) positive.

Table 2 illustrates the same data but for the expert annotator (in total 200 annotations):

Polarity	Number of annotations (Percentage)
Very Negative	32 (16%)
Negative	91 (46%)
Neutral	22 (11%)
Mixed	6 (3%)
Positive	42 (21%)
Very Positive	7 (3%)

Table 2: Frequency distribution of polarity annotations among the expert annotator.

The results show that the distribution of expert annotations is overall quite similar to the annotations of the non-experts, since the majority of annotations are negative (62%). However, one major difference is that the expert annotator rarely used the annotation mixed (3%), while non-experts used it for 23% of all annotations. For binary polarity the distribution of the expert is identical to the non-experts: 134 positive (67%) and 66 negative (33%) annotations. Due to the length constraints of this extended abstract we will not present the results of emotion annotation in detail. However, some major findings are that the most frequent emotion annotations are anticipation (30.6%) and anger (21.1%) while disgust is chosen very rarely (3.9%). We also examined if a speech is annotated with at least one emotion. This is the case for the vast majority of speeches (79.50%).

We performed different statistical tests to analyze the influence of the length of a speech. An analysis of variances with the polarity groups (negative, positive, mixed, neutral) and the length of the speeches shows that there is a significant effect of length on the chosen polarity annotation for non-experts, $F(3, 997)=4.40$, $p=0.004$. Speeches annotated as *mixed* tend to be longer ($M=56.35$, $SD=45.70$) than other speeches and especially than *neutral* annotated speeches, which are on average the shortest type of speeches ($M=41.67$; $SD=28.94$). For the expert annotations, no significant differences among the same polarity groups could be found. However, descriptive analysis also shows that negative ($M=54.52$, $SD=47.31$) and mixed ($M=55.17$, $SD=55.17$) speeches are considerably longer than neutral speeches ($M=32$, $SD=12.98$). We made no significant findings concerning the influence of length on the binary polarity. We also examined statistics concerning the level of agreement (see Table 3). As measures, we chose *Krippendorff's α* (Krippendorff, 2011) and the average percentage of agreement of all annotator pairs (APA).

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

Table 3: Measures of agreement for polarity annotations among non-experts

Krippendorff's α and the APA for the differentiated polarity are very low. However, the level of agreement increases for the binary polarity to a moderate level of agreement. We could not find a significant influence of speech length on the level of agreement. To analyze the difference between the expert and the non-expert annotator we calculated the agreement of every non-expert with the expert separately via Cohen's Kappa (κ) and the APA and then formed the average of these values. (see Table 4).

	Averaged κ values	Averaged APA
Differentiated polarity	0.19	39%
Binary polarity	0.45	76%

Table 4: Averaged measures of agreement for polarity annotations among non-experts with the expert

Krippendorff's α and Cohen's κ are related agreement metrics. Therefore, a comparison is statistically legit. Similar to the agreement solely among non-experts, the level of agreement is very low for differentiated polarity and moderate for the binary polarity. To further analyze differences between the expert and the non-experts for the binary polarity, we compared the annotation value for each speech chosen by the majority of non-experts and compared it to the annotation of the expert. In 43 (21%) cases the expert annotation was different to the annotation of the majority of the non-experts. The numbers are similar when comparing non-experts among each other. With regard to the emotion annotation, the calculation of Krippendorff's α and κ is skewed since the distribution of emotions always shows an excessive proportion of "not present" for all single emotion categories. Therefore, the APA values are rather high ranging from 61% for anticipation to 95% for disgust.

Because of the higher agreement, we chose the binary polarity as the final determinant for the annotation of polarity. We assigned each of the 200 speeches with the consensus of the majority of all annotators (n=6) and whenever there was no majority, the expert annotation was used as a tie-breaker. Therefore a speech is assigned with a category if at least four annotators agree upon it and if it is tied, the annotation of the expert is chosen (this was the case 19 times). As a result, 138 speeches were assigned as

negative and 62 as positive.[1] Table 5 summarizes the results of the questionnaire statements concerning the difficulty of the annotation as well as the confidence about the annotation decisions among the non-experts and the expert.

	Non-experts (n=5)				Expert (n=1)
	Min	Average	Median	Max	
Overall-difficulty	4	5.4	6	6	3
Polarity-difficulty	3	4.6	5	6	3
Emotion-difficulty	3	4.6	5	6	2
Overall-certainty	2	3.4	3	5	4
Polarity-certainty	2	4	4	6	5
Emotion-certainty	2	3.4	3	5	6

Table 5: Descriptive statistics – questionnaire items

A median value of 6 shows that the annotation was perceived as very challenging by the non-experts. With regard to the confidence, the median points to a mediocre certainty for polarities and a rather low certainty for emotion annotations. The expert however reports that she perceived the task to be in-between easy and moderately challenging. However the level of certainty is only slightly higher compared to the mean values of the non-experts. On average participants needed around 5 hours to complete the entire annotation. The expert reported the same amount of time. Analyzing the answers in the free response field as well as the post-annotation interviews, the following major difficulties among non-experts were reported:

- Poetic and archaic language, e.g. unknown words and complex sentences
- Problems in putting a speech in a content-related overall context
- Interpretation of irony and sarcasm
- Multiple Emotions and Polarity-shifts during a speech, especially longer speeches
- Some speeches seem to be meaningless, because they are to short or consist of irrelevant phrases
- The annotation process is perceived as cognitively very challenging; breaks to refocus concentration are needed
- Sometimes the difficulties in understanding the content and context of a speech lead to almost randomly selecting an annotation
- It is not always clear what should be annotated: the sentiment of the language, the sentiment towards a person, the sentiment towards a subject or the emotional state of the speaker?

The feedback of the expert included most of the aforementioned points. However, she didn't report as many difficulties with the language and the context and reported that she often was unsure if she should annotate the sentiment based on the word-level or based on the overall context of the text.

4. Discussion

The overrepresentation of negatively connoted speeches is very dominant and also compliant with findings from Alm and Sproat (2005) in the context of fairy tales. This is a remarkable result, since our specific corpus consists mostly of comedies, which intuitively should be in a rather positive tone (as opposed to tragedies). This overrepresentation is also consistent among the expert and non-experts. We are currently working together with literary scholars to further explore and interpret this phenomenon. The overrepresentation is also an important finding for further annotation studies and sentiment analysis projects, as it suggests an annotation scheme that differs between negative sentiments or that uses continuous scales.

Another finding concerning the distribution of sentiments shows that overall, we have less neutral annotations than in related studies on narrative and historical texts, (Alm & Sproat, 2005; Sprugnoli, 2016). We also found that annotators perceive the presence of at least one emotion for most of the speeches, underlining that dramas are particularly suited and interesting for sentiment analysis. In addition, the class of "mixed" speeches makes for a substantial part of the corpus, at least for the non-experts. According to annotations and the statements of the non-experts, the main reason for this are over-long speeches, which oftentimes contain significant changes of sentiment. Although the expert annotator did not choose the mixed annotation very often, the problem of polarity shifts and multiple emotions was also reported. Overall, we conclude that future annotation schemes should be able to handle such inter-speech changes for a more precise annotation.

As for annotator agreement, we found low to mediocre levels of agreement. This observation is compliant to similar research in the field of narrative and historical texts (Alm & Sproat, 2005; Alm et al., 2005; Sprugnoli et al., 2016), although these studies regard the sentiment of sentences, and not of drama speeches. However, in similar annotation studies with other text sorts much higher levels of agreement are achieved regarding Kappa-statistics, which are comparable to Krippendorff's α. The annotator agreements range from 0.8 to 1.0 for text sorts like movie reviews (Thet et al., 2010), social media comments (Prabowo & Thelwall, 2009), sentences from websites (Kaji & Kitsuregawa, 2007) and microblogs (Bermingham & Smeaton, 2010). For our annotation scenario, it is noticeable that the agreement among non-experts and the agreement between non-experts and the expert annotator are both similarly low to mediocre, i.e. that based on the current data, the difference between an expert and a non-expert is very similar to the difference between two or more non-experts. Overall, the results confirm our assumption that sentiment annotation of narrative texts is more problematic than in other fields. It seems to be a rather subjective annotation task that does not primarily depend on domain expertise. The low agreement is also important for future evaluations of sentiment analysis methods since the level of agreement is often used as performance baseline (Mozetič et al., 2016). We will have to investigate how annotation agreement can be generally improved, e.g.

[1] The corpus with all annotations is available online as a structured table: https://github.com/lauchblatt/LessingSentimentEmotionCorpus (link updated in 09/2020)

by more specific introductions to the task and some common guidelines that give hints on how to use the annotation scheme and how to deal with problems such as uncertainty.

The results of the questionnaire and the concluding interview support these claims. Non-expert participants perceive the annotation as very difficult and they report to have mediocre certainty about the correctness of their annotation. They state that the task is cognitively very challenging and that it demands high levels of concentration throughout the process. Non-experts also had issues with the historic language and the context of some speeches. In contrast, the expert did not perceive the annotation task too difficult and demanding, and also only reported minor issues with language and missing context.

The low agreements of more complex schemes would suggest the usage of a rather simple scheme (e.g. binary polarity). However, the results of the interviews also show that the annotation schemes derived from application areas like product reviews might not be suitable for the use case of literary text. For example, annotators did not know how to mark irony, sarcasm or multiple polarity shifts. The annotators also noted that there are often multiple possible targets for the annotation of a sentiment and that it is not always clear which sentiment to choose. Based on this feedback we suggest to extend the scheme so that the annotators can distinguish the reference of the sentiment, e.g. another speaker, a topic or speaker that is directly or indirectly talked about, etc. (cf. Shin et al., 2012). As for another challenge, some annotators also mentioned that they were sometimes inclined to interpret sentiment from a rather subjective perspective, as they had personal associations with some of the speeches. Future research should pay attention to these problems and instructions and annotation schemes should be as clear and precise as possible to avoid confusion.

One of our main goals was to explore if non-experts are potentially capable to perform sentiment annotation for historic plays, because non-experts are obviously more available than experts, which is an important aspect for the design of future large-scale studies. We found that non-experts perceived the task as more challenging and more grave problems occurred, e.g. not understanding the language or the context correctly. While the usage of non-experts in the annotation process is not uncommon for sentiment analysis (Volkova et al., 2010), we found that they seemed to struggle in our particular annotation scenario. On the other side, the agreement between non-experts and the expert is not any different than among non-experts only, which indicates that experts also struggle with the task. This observation is reflected by related studies of Alm and Sproat (2005) as well as Sprugnoli et al. (2016), who also report low levels of agreement while using trained students or even more advanced experts. Assuming expertise is an important factor, the agreement between a non-expert and the expert should be notably lower. The distribution of polarities is also very similar between non-experts and the expert. Further, taking the majority decision of all non-expert annotators leads to annotations that are very similar to the expert's annotation. With regard to the time needed to achieve the complete annotation task, there are also no major differences to be found, as it took around 5 hours to finish the annotation for both, the expert and the non-experts.

5. Conclusion and Future Directions

We believe it is feasible to use non-experts in a large-scale crowdsourcing context for the annotation of historic plays, keeping in mind the improvements regarding the annotation scheme and instructions mentioned before. As for the language problems, non-expert annotators could be provided with a lexicon of the most frequent historic words. This lexicon could also be used to filter speeches that contain problematic language, which could then be reserved for an expert annotator. The issues of missing context could be easily resolved by providing a digital tool for the annotation task (which would be needed for a large-scale study in any chase), which would allow for the optional display of arbitrary portions of context.

As our annotation study was solely focused on speeches, more complex structural levels such as scenes, acts, speakers or speaker relations could also be taken into account for future studies. The interpretation of these levels would also be necessary to get a more complete view on a drama. Another problem is that feedback by both, the expert and the non-experts, points to the lack of precise instructions to the sentiment annotation task, which certainly is an influencing factor for low agreements.

Furthermore, we are aware that our sample size with only one expert is very small, so further research will be necessary to explore which level of expertise is tolerable and if there are also significant differences in annotation behavior between experts and non-experts in a larger study. We are currently conducting a follow-up annotation study with trained students of German literary studies to analyze if the problems described in this article persist, if differences in the annotation process occur and if their level of expertise is sufficient. For this study, we will adjust our annotation scheme and use a bigger corpus and more participants.

We also want to further examine how literary scholars annotate sentiment and which requirements an annotation scheme for this context has to meet. As a long-term goal, we would like to develop an annotation scheme optimized for the context of drama sentiment annotation. By this, we hope be able to develop tools for more efficient sentiment annotation and to acquire large-scale annotated corpora for evaluation and machine learning purposes.

6. Bibliographical References

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.

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.

Bermingham, A., & Smeaton, A. F. (2010). Classifying sentiment in microblogs: is brevity an advantage?. In

Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1833-1836). ACM.

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).

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.

Kaji, N., & Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).*

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

Kao, J., & Jurafsky, D. (2012). A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature* (pp. 8-17).

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

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

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

Meyer-Sickendiek, B. (2005). *Affektpoetik: eine Kulturgeschichte literarischer Emotionen.* Würzburg: Königshausen & Neumann.

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.

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 classifica-tion: 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).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.

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

Refaee, E., & Rieser, V. (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In *LREC* (pp. 2268-2273).

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In: *1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI* (ESSEM 2013).

Schmidt, T., Burghardt, M. & Dennerlein, K. (2018). "Kann man denn auch nicht lachend sehr ernsthaft sein?" – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen. In *Book of Abstracts*, DHd 2018.

Shin, H., Kim, M., Jang, H., & Cattle, A. (2012). Annotation Scheme for Constructing Sentiment Corpus in Korean. In *PACLIC* (pp. 181-190).

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.

Volkova, E. P., Mohler, B. J., Meurers, D., Gerdemann, D., & Bülthoff, H. H. (2010). Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 98-106). Association for Computational Linguistics.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2), 165-210.

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.