Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's *Emilia Galotti*

Thomas Schmidt [1], Manuel Burghardt [2], Christian Wolff [1]

[1] Media Informatics Group, University of Regensburg, 93040, Regensburg, Germany
[2] Computational Humanities Department, University of Leipzig, 04109 Leipzig, Germany
thomas.schmidt@ur.de
burghardt@informatik.uni-leipzig.de
christian.wolff@ur.de

Abstract. We present a case study as part of a work-in-progress project about multimodal sentiment analysis on historic German plays, taking *Emilia Galotti* by G. E. Lessing as our initial use case. We analyze the textual version and an audio version (audiobook). We focus on ready-to-use sentiment analysis methods: For the textual component, we implement a naive lexicon-based approach and another approach that enhances the lexicon by means of several NLP methods. For the audio analysis, we use the free version of the *Vokaturi* tool. We compare the results of all approaches and evaluate them against the annotations of a human expert, which serves as a gold standard. For our use case, we can show that audio and text sentiment analysis behave very differently: textual sentiment analysis tends to predict sentiment as rather negative and audio sentiment as rather positive. Compared to the gold standard, the textual sentiment analysis achieves accuracies of 56% while the accuracy for audio sentiment analysis is only 32%. We discuss possible reasons for these mediocre results and give an outlook on further steps we want to pursue in the context of multimodal sentiment analysis on historic plays.

Keywords: sentiment analysis, emotion analysis, multimodal, multimedia, computational literary studies, audio, audiobooks, drama, text mining, Lessing

1 Introduction

Sentiment analysis is the area of research that deals with computational methods to analyze and predict sentiments and emotions in written text (Liu, 2016, p.1). Although the majority of work in this area is done with user-generated content like product reviews and social media (Vinodhini & Chandrasekaran, 2012), there is growing interest in exploring the application of sentiment analysis in the digital humanities especially in computational literary studies. Sentiment analysis is used to analyze fairy tales, (Alm et al., 2005; Mohammad, 2011) novels, (Kakkonen & Kakkonen, 2011; Jockers, 2015) historic plays, (Mohammad, 2011; Nalisnick & Baird, 2013) or to generate features for various machine-learning tasks (Jannidis et al., 2016; Kim, Padó & Klinger, 2017).

Sentiment analysis as a whole and especially the current research in the digital humanities is predominately focused on the analysis of written text; other media channels like audio, video or combinations are neglected so far. We propose several reasons for justifying the broadening of the current focus on written text to other modalities: First, although systematic performance evaluation in the context of narrative texts is rare, the existing research shows that when compared to human sentiment annotations, the accuracy can vary between 20-70% depending on the method and the type of text (Kim & Klinger, 2018; Schmidt & Burghardt, 2018a; Schmidt & Burghardt 2018b). Therefore, it is far lower than in other areas of sentiment analysis in which accuracies close to and above 90% are achieved (Vinodhini & Chandrasekaran, 2012). Multimodal sentiment analysis has been proven to be more successful than isolated text sentiment analysis in several areas (Morency et al., 2011; Abburi et al., 2016; Poria et al., 2017) and one can observe a general trend from unimodal to multimodal sentiment analysis approaches (Poria et al., 2017). We hypothesize that narrative text might be especially suited for multimodal sentiment analysis. One reason for the current low accuracies might be that prosodic and voice-based features, which are important parts of the narrative performance and the expression of emotions, are neglected in text sentiment analysis. Especially in the context of plays, which are specifically designed for oral performance in a theatre, the voice and the face of actors are important identifiers for emotion and sentiment. Furthermore, audio and video/face sentiment analysis are far less language dependent than the current text-based approaches (Hudlicka, 2003) which is especially appealing for research in multilingual literary studies.

Second, we also see potential for using differing media channels to improve the annotation of sentiment for narrative texts. Literary texts have been proven to be very difficult and tedious to annotate due to the historic and complex language (Schmidt, Burghardt & Dennerlein, 2018; Schmidt, Burghardt & Wolff, 2018; Alm et al., 2005). The presentation of text material in multimodal form might improve and facilitate the annotation process concerning sentiment. For example, annotators might not understand the language and the context of a narrative text unit but the expressed emotion of an actor in his oral performance. Improvement in the annotation process would enable us to acquire annotated corpora for evaluation and machine learning purposes on a larger scale more easily.

In this paper, we contribute to "multimodality in digital humanities" by presenting first work-in-progress results for multimodal sentiment analysis on narrative texts, more precisely for the specific use case of the play *Emilia Galotti* by G. E. Lessing. We have analyzed and compared existing text sentiment analysis approaches and a ready-to-use audio speech sentiment analysis. In addition, we have evaluated the performance compared to a gold standard of annotations by a human expert. Finally, we discuss the results and the limitations of this case study but also formulate an agenda for future research in this area.

2 Research Questions

Sentiment analysis and emotion analysis are often differentiated in research. While sentiment analysis is regarded as predicting and analyzing the overall affective state predominantly with the classes *positive*, *negative* and *neutral* (we refer to these classes as polarity), emotion analysis deals with more complex emotional categories like anger, surprise or joy (Vinodhini & Chandrasekaran, 2012). For this case study, we focus solely on sentiment since its application is in general easier and more successful (Liu, 2016, p.67) especially in the context of literary texts (Alm et al., 2005).

As modalities for this study, we use the textual and an audio version (audiobook) of Emilia Galotti. We want to explore the following research questions:

> *RQ1: How do text sentiment analysis approaches perform compared to a ready-to-use audio sentiment analysis approach?*

We are focusing on simple and ready-to-use solutions, since not only is this study our first exploration in this area, but also because it is a common use case in digital humanities when the focus of research is not the development of new algorithms but the analysis of cultural artifacts. With RQ1, we want to gather first insights in the possible similarities and differences of both media channels and analysis types.

> *RQ2: How do text sentiment analysis approaches and the audio sentiment analysis perform against human expert annotations (with text)?*

We want to use the sentiment annotations of an expert as gold standard and evaluate the text and audio approach against each other. In our case study, the expert annotated the sentiment by being presented with the text. With RQ2, we want to test our assumption that audio speech conveys more precise emotional information and that audio sentiment analysis therefore achieves higher accuracies.

3 Methods and Data

As use case for our investigations, we chose the play *Emilia Galotti* by G. E. Lessing (premiered 1772). The reason for this is that our recent research is focused on Lessing and *Emilia Galotti* is one of his most famous plays, which means that audio material for this play is available, too. All analyses are speech based, thus we are only comparing singular speeches with each other. A speech is a single utterance of a character separated by utterances of other characters beforehand and afterwards. Overall, the play consists of 835 speeches. The longest speech consists of 235 words while the shortest speech only has one word. On average, a speech in this play consists of 23 words.

As material for the textual sentiment analysis, we gathered an XML-annotated version of the play from the platform *Textgrid*[1]. For the sentiment analysis, we employ two lexicon-based approaches. A sentiment lexicon is a list of words annotated with sentiment annotation. Based on simple calculation, text units can be assigned a polarity (e.g. neutral, positive, negative). In previous work (Schmidt & Burghardt, 2018a, Schmidt & Burghardt, 2018b) we have evaluated different lexicons and NLP approaches for the

[1] https://textgrid.de/ (Note: all URLs mentioned in this article were last checked Feb. 10, 2019)

best performance on a corpus of all of Lessing's plays. For the use case in this paper we analyze (1) the existing German sentiment lexicon *SentiWS* (Remus et al., 2010) as is and (2) a variant where we enhance SentiWS with additional NLP techniques and preprocessing methods. The enhanced SentiWS approach achieved the highest accuracy in a previous study on a Lessing corpus. In this approach lemmatization as well as the extension of the lexicon with historical variants are employed (for more details see Schmidt & Burghardt, 2018a). We refer to the first approach as naive (lexicon-based) approach and to the second one as optimized (still lexicon-based) approach. Both approaches produce numerical values with values below zero being assigned as negative sentiment, over zero as positive and equal to zero as neutral.

For the audio analysis, we at first screened different available commercial and non-commercial audiobooks. However, we identified several problems: Those audiobooks use different speakers and a lot of additional sound effects and background music, which might be problematic for the sentiment analysis without larger preprocessing steps. Furthermore, audiobooks often differ strongly from the original text, leaving whole passages out or switching the order of speeches. We found the best material for this use case via semiprofessional readings of the original text on several platforms on the web. We used publicly available recordings from YouTube[2]. The reader is female and there are no deviations from the original text in the reading. Several piano pieces are included in the recording, but they are separate from the general reading.

For the audio sentiment analysis, we use the free version of *Vokaturi*[3]. *Vokaturi* is an emotion recognition software for spoken language with an easy-to-use API. Garcia-Garcia et al. (2017) recommend *Vokaturi* as the best free software for spoken language sentiment analysis. *Vokaturi* is described as being language independent and it works via machine learning with two annotated databases. *Vokaturi* takes audio data of any length and outputs five values that range from 0 (none) – 1 (a lot) for the categories *neutrality*, *fear*, *sadness*, *anger* and *happiness*.

The workflow for the audio sentiment analysis is as follows: In a preprocessing step we trimmed and transformed the audio files from YouTube. We then performed forced alignment with the free Python library *aeneas*[4]. Forced alignment is a method to align text segments and audio speech and determine precise time stamps of when the text segments in the audio file start and end. As text segments we used the 835 speeches of the play. According to the time stamps, we segmented the audio file to get 835 separated audio files, which are finally used with *Vokaturi*. To map the output of *Vokaturi* to the nominal scale used for the textual sentiment analysis we employ a heuristic mathematical approach. First, we sum up all values for the negative emotions fear, sadness and anger to get a value for negative sentiment. We regard the value for happiness as positive sentiment. We then chose the maximum value of negative sentiment, positive sentiment and the value for neutrality as the final polarity of a speech.

[2] The entire playlist of the files are available online: https://www.youtube.com/playlist?list=PL06w7wmahre2eMZXDBgGcR2mgAoeI_f8k

[3] Available online: https://developers.vokaturi.com/getting-started/overview

[4] https://github.com/readbeyond/aeneas

The human annotations we use were gathered with an expert literary scholar who annotated 200 random speeches of Emilia Galotti. The annotator was presented the speech to be annotated, the predecessor and successor speech for context and then had to annotate the polarity as rather positive, neutral or rather negative. The annotator was instructed to annotate the polarity he feels is most connoted with the speech. Note that the speeches were solely presented in textual form. The annotator was a male student of German literary studies who had to write a thesis about Emilia Galotti during the annotation process and can therefore be regarded as an expert for this specific play. We restricted the annotation to 200 speeches since sentiment annotation in this context has been proven to be very tedious and challenging (Alm et al., 2005; Schmidt, Burghardt & Dennerlein, 2018). Therefore, all comparisons with human annotations are done with those specific 200 speeches. More information about the annotation process can be found in Schmidt, Burghardt and Dennerlein (2018), where we performed a very similar annotation study.

4 Results

We first report all results concerning the comparison of the text and audio sentiment analysis among all 835 speeches (RQ1). Table 1 shows the overall polarity distribution of all methods.

Table 1. Polarity distributions among all methods (naive = SentiWS, optimized = SentiWS + NLP, Audio = Vokaturi).

	Text: Naive	*Text: Optimized*	*Audio*
negative	215 (26%)	411 (49%)	289 (35%)
neutral	340 (40%)	198 (24%)	121 (14%)
positive	280 (34%)	226 (27%)	425 (51%)

Table 2 is a cross table that illustrates the comparison of the distributions of polarity classes between the naive lexicon and the audio-based approach. The number of similar assigned speeches per class is in bold.

Table 2. Cross table for naive lexicon-based (SentiWS) and audio (Vokaturi) approach.

		Audio			
	Polarity	*negative*	*neutral*	*positive*	*Sum*
Text: Naive	*negative*	**54**	26	135	215
	neutral	154	**54**	132	340
	positive	81	41	**158**	280
	sum	289	121	425	835

The majority of the assignments of the naive approach are neutral while the majority (51%) of all speeches are assigned as positive by the audio sentiment analysis (see Table 1). Therefore the proportion of similarly assigned speeches is rather low (32%); 266

speeches are assigned the same class. Table 3 shows the same cross table but with the optimized lexicon-based approach.

Table 3. Cross table for optimized lexicon-based (SentiWS+NLP) and audio (Vokaturi) approach.

		Audio			
	Polarity	*negative*	*neutral*	*positive*	*Sum*
Text: Optimized	*negative*	**109**	49	253	411
	neutral	108	**31**	59	198
	positive	72	41	**113**	226
	sum	289	121	425	835

The optimized lexicon approach predicts that almost half of the speeches are negative while the audio sentiment analysis approach produces opposite results with half of the speeches being assigned as positive (51%; see Table 1). Consequently, a small number of 253 speeches are assigned the same class. This results in 30%.

For RQ2 we regard the 200 human annotated speeches and the performance of all approaches compared to the human annotations. First, Table 4 shows the polarity distributions of all methods on this subset of speeches.

Table 4. Sentiment distributions for all computational methods and the expert annotation

	Text: Naive	*Text: Optimized*	*Audio*	*Expert text annotation*
negative	58 (29%)	97 (48%)	59 (29%)	92 (46%)
neutral	82 (41%)	53 (27%)	29 (15%)	60 (30%)
positive	60 (30%)	50 (25%)	112 (56%)	48 (24%)

The polarity distributions for the different computational methods on the subset of the annotated corpus are similar to the distributions on the entire corpus. Most of the speeches were annotated as negative by the expert annotator (46%). With the following cross tables, we show the agreements of the computational methods and the human annotations. The human annotations are used as gold standard to evaluate the approaches. The accuracy is the proportion of correctly predicted speeches among all speeches. First, Table 5 and 6 show the text-based sentiment analysis results.

Table 5. Cross table for the expert annotation and the naive (SentiWS) lexicon based approach

		Text: naive			
	Polarity	*negative*	*neutral*	*positive*	*Sum*
Expert text annotation	negative	**46**	21	25	92
	neutral	3	**40**	17	60
	positive	9	21	**18**	48
	Sum	58	82	60	200

Table 6. Cross table for the expert annotation and the optimized (SentiWS+NLP) lexicon based approach

		Text: optimized			
	Polarity	*negative*	*neutral*	*positive*	*Sum*
Expert text annotation	negative	**65**	13	14	92
	neutral	15	**28**	17	60
	positive	17	12	**19**	48
	Sum	97	53	50	200

Taking the human annotation as benchmark, both textual approaches perform almost similarly: The naive approach achieves an accuracy of 52% (104 speeches) and the optimized approach 56 % (112 speeches). Both approaches are over the random baseline (approx. 36%) and slightly above the majority baseline (46%).

Table 7. Cross table for the expert annotation and the audio (Vokaturi) approach

		Audio			
	Polarity	*negative*	*neutral*	*positive*	*Sum*
Expert text annotation	*negative*	**22**	13	57	92
	neutral	23	**11**	26	60
	positive	14	5	**29**	48
	Sum	59	29	112	200

With the human annotation as gold standard, the audio sentiment analysis achieves an accuracy of 31% with 62 speeches being predicted correctly. This is below the random baseline. The main difference is that the human annotator chose negative annotations for most of the speeches while the audio sentiment analysis in contrast predicts positive sentiment for the majority of times.

5 Discussion

With the first research questions, we analyzed and compared basic text and audio sentiment analysis approaches. We identified that the sentiment analysis on both, text and audio, produces very different results with the optimized text analysis predicting the majority of speeches as negative and the audio speech sentiment analysis in contrast predicting the majority as positive, more specifically as connoted with the emotion happiness. For our specific use case, we were able to show that both channels seem to work on very different levels and with very different features, which exemplifies that the inclusion of other media channels leads to new insights, in this case even contradictory results.

We assume that there might be some specifics of the audio material that lead to this tendency of positive sentiment assignments. Firstly, pitch plays an important factor in the prediction process (Garcia-Garcia et al., 2017). We analyzed several examples of speeches falsely assigned as positive and noticed that the reader uses a high pitch voice for effect reasons in several instances. Furthermore, our speaker is female and therefore

has a general higher pitch. We assume that these factors might falsely direct the algorithm to the assignment of higher happiness levels. Additionally, note that we used a heuristic approach to transform the emotion categories of Vokaturi into sentiment. However, in several research areas emotion and sentiment are regarded as rather different concepts (Liu, 2016, p.31-39) and a transformation like this might be too simplistic.

When evaluating the computational approaches against a gold standard of a human annotated subset of 200 speeches the general problems and limitations of sentiment analysis on literary texts are apparent. The textual approaches as well as the audio-based method perform rather poorly: the textual approach is just slightly above the majority baseline, the audio approach is below the random baseline. This is in line with current research on sentiment analysis on literary texts (Schmidt & Burghardt, 2018a; Kim & Klinger, 2018). Historic narrative texts continue to emerge as a very challenging text sort for sentiment analysis.

Our assumption that the audio speech analysis might improve the results has been proven wrong on the chosen material. The text sentiment analysis performs far better on the gold standard. The reason for this is predominately that the human annotator as well as the text analysis assign the majority of speeches as negative while the audio sentiment analysis behaves contrarily.

To put these results in perspective one should also note that audio sentiment analysis is in general known as being more challenging than other media channels like text and facial expressions via video and performs often lower than those (Hudlicka, 2003; Poria et al., 2017). Also, note that there are far more ready-to-use software and APIs for textual and facial sentiment analysis than there are for audio (Poria et al., 2017). Audio sentiment analysis on standardized corpora achieves accuracies up to 75% (Poria et al., 2017). However, in our use case the audio sentiment analysis performs far lower. Note that the annotator used just the text for annotation and did not annotate or use the audio speech, which also might be a reason that the text sentiment analysis and the human annotation are more in line with each other. For example, the interpretation and oral performance of the reader might have been more positive than the text itself implies.

In future work, we want to investigate how the audio speech influences the human annotation. Furthermore, bear in mind that the performance comparison is not the main goal of our efforts. In future research, we rather want to explore possibilities of combining multiple media channels to improve performance results.

Overall, there are several limitations in this case study we want to address in future research. The presented study represents only a singular use case and a work-in-progress project. Onwards, we want to analyze more examples like different audio material of one play and in general plays of other writers and eras. We also want to explore different audio sentiment analysis approaches. The usage of ready-to-use solutions like *Vokaturi* in general sentiment analysis research is rather rare; usually an individual machine-learning algorithm is implemented. For this specific reason, we plan several annotation studies to acquire annotated audio material on a large scale. In this context, we also want to explore how the audio or video presentation of speeches can improve the sentiment annotation process in this area.

Furthermore, we also want to include the media channel of video especially facial emotion recognition in our research via video recordings of theatrical performances of

the plays. The inclusion of the oral but also facial expression of an actor provide necessary interpretation channels for a holistic understanding of the sentiment and emotion in a play. With these future plans, we want to continue our work towards a multimodal sentiment analysis and explore possibilities for the technical performance, the annotation process and the interpretation in literary studies.

References

Abburi, H., Akkireddy, E. S. A., Gangashetti, S., & Mamidi, R. (2016). Multimodal Sentiment Analysis of Telugu Songs. In *SAAIP@ IJCAI* (pp. 48-52).

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).

Garcia-Garcia, J. M., Penichet, V. M., & Lozano, M. D. (2017). Emotion detection: a technology review. In *Proceedings of the XVIII International Conference on Human Computer Interaction* (p. 8).

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *International journal of human-computer studies*, 59(1-2), 1-32.

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). *Revealing sentiment and plot arcs with the syuzhet package*. Retrieved from http://www.matthewjockers.net/2015/02/02/syuzhet/

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).

Kim, E. & Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In: *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1345-1359). Association for Computational Linguistics.

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

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

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).

Morency, L. P., Mihalcea, R., & Doshi, P. (2011, November). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 169-176). ACM.

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).

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98-125.

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

Schmidt, T. & Burghardt, M. (2018a). An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In: *SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH-CLfL 2018)* (pp. 139-149). Retrieved from http://aclweb.org/anthology/W18-4516

Schmidt, T. & Burghardt, M. (2018b). Toward a Tool for Sentiment Analysis for German Historic Plays. In: Piotrowski, M. (ed.), *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018* (pp. 46-48). Lausanne, Switzerland: Laboratoire laussannois d'informatique et statistique textuelle. Retrieved from https://zenodo.org/record/1312779#.XGA0iFxKhPZ

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. Retrieved from http://ceur-ws.org/Vol-2155/schmidt.pdf

Schmidt, T., Burghardt, M. & Wolff, C., (2018). Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten. In: Burghardt, M. & Müller-Birn, C. (Hrsg.), *INF-DH-2018*. Bonn: Gesellschaft für Informatik e.V. Retrieved from https://dl.gi.de/handle/20.500.12116/16996

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.