Sentiment Annotation for Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts

Schmidt, Thomas
Media Informatics Group, University of Regensburg, Germany
http://mi.ur.de/sekretariat-team/thomas-schmidt
thomas.schmidt@ur.de

Burghardt, Manuel
Computational Humanities, University of Leipzig, Germany
https://ch.uni-leipzig.de/
burghardt@informatik.uni-leipzig.de

Dennerlein, Katrin
University of Würzburg, Germany
https://www.germanistik.uni-wuerzburg.de/ndl1/mitarbeiter/dennerlein/
katrin.dennerlein@uni-wuerzburg.de

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

Abstract

We present first results of an ongoing research project on sentiment annotation of historical plays by German playwright G. E. Lessing (1729-1781). For a subset of speeches from six of his most famous plays, we gathered sentiment annotations by two independent annotators for each play. The annotators were nine students from a Master's program of German Literature. Overall, we gathered annotations for 1,183 speeches. We report sentiment distributions and agreement metrics and put the results in the context of current research. A preliminary version of the annotated corpus of speeches is publicly available online and can be used for further investigations, evaluations and computational sentiment analysis approaches.

1 Introduction

Sentiment analysis typically makes use of computational methods to detect and analyze sentiment (e.g. *positive* or *negative*) in written text [7, p.1]. Emotion analysis, a research area very closely related to sentiment analysis, deals with the analysis of more complex emotion categories like anger, sadness or joy in written text. More recently, sentiment and emotion analysis have also gained attention in computational literary studies [15] and are used to investigate novels [3, 4], fairy tales [1, 8] and plays [8, 9, 11, 12, 13]. However, there are only few resources available for sentiment analysis on literary texts, since sentiment annotation of literary texts has turned out to be a tedious, time-consuming, and generally challenging task [5, 14, 15], especially for historical literary texts, with their archaic language and complex plots [14]. Previous studies have shown that agreement levels among annotators of sentiment annotation are rather low, since literary texts can be understood and interpreted in a number

of ways, i.e. annotations are very much dependent on the subjective understanding of the annotator [14, 15]. On the technical side, most computational approaches for sentiment analysis on literary texts currently employ heuristic and rule-based methods [4, 8, 9, 11, 12, 13]. For more advanced machine learning tasks as well as for standardized evaluations, well-curated sentiment-annotated corpora for literary texts would be an important desideratum, but in practice they are largely missing.

We want to close this gap for the genre of German historical plays, more precisely plays of the playwright *G. E. Lessing* (1729-1781), by creating a corpus annotated with sentiment information. Since this is a very tedious task, we first want to investigate the suitability of different user groups with different levels of knowledge for literary works. We want to find out how important domain knowledge (*experts, semi-experts* and *non-experts*) is for this kind of task, since the acquisition of domain experts (literary scholars) is obviously more difficult than the acquisition of annotators without any formal literary training, who could possibly be hired on a large scale via a crowdsourcing platform. In this article we present first results from an annotation study with a group of semi-experts.

2   Design of the Annotation Study

We conducted the annotation study for semi-experts as part of a course in the Master's program of German Literature at the Würzburg University. The course *"Sentiment Analysis vs. Affektlehre"* was focused on the plays of Lessing and revolved around the comparison of more traditional sentiment and emotion analysis approaches with recent computational methods [11, 12]. Each student had to write an essay and a presentation for a specific play. In addition, students were asked to participate in the sentiment annotation study. For this study, we chose six of the most famous plays by Lessing. Every student had to annotate 200 randomly selected speeches except for the students assigned with *Damon*, since this play only consists of 183 speeches. A speech is a single utterance of a character, typically separated by utterances of other characters beforehand and afterwards. A speech can consist of one or multiple sentences. For each of the six plays we asked two independent students to provide sentiment annotations. Overall, nine students participated in the annotation project. One student, who was an advanced tutor, conducted annotations for multiple plays. The entire corpus comprises 1,183 speeches and we acquired 2,366 annotations (2 annotations per speech). The corpus consists of 3,738 sentences and 44,192 tokens. A speech consists on average of 3.16 sentences and 37.36 tokens.

The overall process was very similar to [14]: Students received an annotation instruction and guidelines of three pages length, explaining the entire annotation process with examples. The annotation material was provided via *Microsoft Word*, which was also used as a basic annotation tool during the annotation process. Conducting annotation studies with well known products like *Word* or *Excel* is not uncommon in Digital Humanities projects, as humanities scholars are typically familiar with standard software tools [2, 14]. Every student received a file with the play-specific speeches. A speech was presented with the following information: (1) speaker name, (2) text of the speech and (3) position of the speech in the entire play. Furthermore, the predecessor and successor speech were presented to give some context information for the interpretation of the actual speech. Sentiment annotations were documented in a predefined table structure. Annotators were asked to write down the overall sentiment of an entire speech rather than the sentiment for specific parts of that speech. In case there were multiple, possibly conflicting sentiment indicators in one speech, annotators documented the most dominant sentiment for that speech.

Students were asked to annotate one of the following classes in the first table: *negative, positive, neutral, mixed, uncertain*, and *other* (we refer to this scheme as *differentiated polarity*). If annotators did not choose negative or positive, they were asked to provide a tendency for one of those two classes (*binary polarity*). Figure 1 illustrates the annotation process via an example annotation. Here the annotator chose *neutral* for the differentiated polarity and *negative* for the binary polarity.

Emilia Galotti 1.Act, 4.Scene, 3.Speech, Number in the entire play: 27

DER PRINZ:

Vortrefflich, Conti; – ganz vortrefflich! – Das gilt Ihrer Kunst, Ihrem Pinsel. – Aber geschmeichelt, Conti; ganz unendlich geschmeichelt!
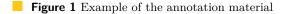
CONTI:

Das Original schien dieser Meinung nicht zu sein. Auch ist es in der Tat nicht mehr geschmeichelt, als die Kunst schmeicheln muß. Die Kunst muß malen, wie sich die plastische Natur, – wenn es eine gibt – das Bild dachte: ohne den Abfall, welchen der widerstrebende Stoff unvermeidlich macht; ohne das Verderb, mit welchem die Zeit dagegen an kämpft.

DER PRINZ:

Der denkende Künstler ist noch eins so viel wert. – Aber das Original, sagen Sie, fand dem ungeachtet –

Differentiated polarity					
Negative	Positive	Neutral	Mixed	Uncertain	Other
		x			

Binary polarity	
Negative	Positiv
x	

Figure 1 Example of the annotation material

We also gathered annotations about the reference of the sentiment and the certainty of the annotations to get more sophisticated insights about the annotation behavior. However, we will only focus on the polarity in the upcoming sections. Students had a month to complete the annotations, but other than that could freely organize the time to complete the task by themselves. In the next section, we report some of the major findings from our annotation study.

3　Results

First, we present the distributions concerning the differentiated polarity for the entire corpus and all 2,366 annotations (table 1):

Negative	Positive	Neutral	Mixed	Uncertain	Other
734 (31%)	442 (19%)	540 (23%)	405 (17%)	235 (9%)	10 (0.4%)

Table 1 Sentiment distribution for all annotations

Most of the annotations are negative, and a significant number of annotations are mixed and uncertain. These results are in line with current research about sentiment annotation of

literary texts [1, 14], which seem to have a general tendency toward more negative annotations. To analyze the agreement among annotators we used *Cohen's Kappa* and the average observed agreement(AOA, number of agreements divided by the total number of annotated speeches) as basic metrics. According to [6], Kappa metrics can be interpreted like this (table 2):

Kappa value	Interpretation
<0.20	Poor agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Very good agreement

■ **Table 2** Interpretation of Cohen's Kappa

As described above, we had the annotators document differentiated (6 values) as well as binary (2 values) polarity values for each of the speeches. In addition, we also inferred a third variant from those annotations, which we call *threefold polarity*. We basically extend the binary polarity metric by the value "neutral" in the following manner: if the value "neutral" is annotated at the differentiated polarity, this value is taken over in this variant. For any other case the value "positive" or "negative" is taken over from the binary polarity annotation. We present these metrics per play along with average values for the entire corpus (table 3):

Play	Annotation type	kappa	AOA
	Differentiated polarity	0.03	15%
Damon	Threefold polarity	0.12	45&
	Binary polarity	0.14	56%
	Differentiated polarity	0.43	59%
Emilia Galotti	Threefold polarity	0.49	68%
	Binary polarity	0.49	76%
	Differentiated polarity	0.13	30%
Der Freigeist	Threefold	0.28	57%
	Binary polarity	0.34	68%
	Differentiated polarity	0.29	44%
Minna von Barnhelm	Threefold polarity	0.39	59%
	Binary polarity	0.29	63%
	Differentiated polarity	0.32	46%
Nathan der Weise	Threefold polarity	0.40	62%
	Binary polarity	0.43	72%
	Differentiated polarity	0.38	54%
Miß Sara Sampson	Threefold polarity	0.48	66%
	Binary polarity	0.65	83%
	Differentiated polarity	0.30	45%
Overall average	Threefold polarity	0.36	59%
	Binary polarity	0.39	69%

■ **Table 3** Agreement levels

4   Discussion

The agreement levels are poor to fair for many annotation types and plays (0.00-0.40). For annotation types with fewer classes and for several plays we identified Kappa values with moderate agreement (0.41-0.60). The highest agreement levels are achieved for *Emilia Galotti* and *Miss Sara Sampson*, in the latter case with substantial agreement for the binary annotation (0.61-0.80). Overall, these results are in line with current research [1, 5, 14], proving that sentiment annotation on literary texts is very subjective, difficult and dependent on the interpretation of the annotator, thus leading to rather low agreement levels compared to sentiment annotations on other text sorts like movie reviews [17] and social media content [10]. However, for several plays and annotation types higher agreements were achieved than compared to similar annotations with non-experts [14]. Therefore, we propose to explore sentiment annotations with advanced experts concerning the literary text to acquire more stable annotations. Although the agreement levels are too low to use the full corpus for evaluation or machine learning purposes, we still think that the corpus can be useful for the computational literary studies community and therefore provide a first alpha version of the corpus with all annotations[1].

5   Future Directions

As a next step, we want to gather more annotations by other user groups and compare them to each other. Although non-experts currently seem to produce lower levels of agreement, we want to explore if we can produce valuable corpora with non-experts by using majority decision with a high number of annotations per speech. To gain insights for possible improvements concerning the annotation scheme and process, we also conducted a focus group with the annotators and acquired feedback via a questionnaire. Note that we focused on the overall sentiment of an entire speech. In the future we want to use more sophisticated models including more precise annotations of the sentiment direction and object. Taking into account existing studies on multi-modal sentiment analysis on dramatic texts [16], we currently also explore the possibility to extend the speech annotation to modalities other than text, e.g. audio and video material of the corresponding theater plays to improve agreement levels.

───── **References** ─────

1   Cecilia Ovesdotter Alm and Richard Sproat. Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction*, pages 668–674. Springer, 2005.

2   Lars Döhling and Manuel Burghardt. PaLaFra – Entwicklung einer Annotationsumgebung für ein diachrones Korpus spätlateinischer und altfranzösischer Texte. In *Book of Abstracts, DHd 2017*, Bern, Switzerland, 2017.

3   Fotis Jannidis, Isabella Reger, Albin Zehe, Martin Becker, Lena Hettinger, and Andreas Hotho. Analyzing features for the detection of happy endings in german novels. *arXiv preprint arXiv:1611.09028*, 2016.

4   Tuomo Kakkonen and Gordana Galic Kakkonen. Sentiprofiler: Creating comparable visual profiles of sentimental content in texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 62–69, 2011.

───────────────

[1]   A preliminary version of this resource is available online:
https://www.dropbox.com/sh/8mu29ny8fhrpgg2/AABFXw7qYHLoJ-4yx8CBlXX9a?dl=0

5   Evgeny Kim and Roman Klinger. 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*, pages 1345–1359, 2018.

6   Rainer Leonhart. *Lehrbuch Statistik: Einstieg und Vertiefung*. Verlag Hans Huber, 2013.

7   Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2016.

8   Saif Mohammad. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics, 2011.

9   Eric T Nalisnick and Henry S Baird. Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 479–483, 2013.

10 Rudy Prabowo and Mike Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, 2009.

11 Thomas Schmidt and Manuel Burghardt. An evaluation of lexicon-based sentiment analysis techniques for the plays of gotthold ephraim lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149. Association for Computational Linguistics, 2018. URL: `http://aclweb.org/anthology/W18-4516`.

12 Thomas Schmidt and Manuel Burghardt. Toward a tool for sentiment analysis for german historic plays. In Michael Piotrowski, editor, *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018*, pages 46–48, Lausanne, Switzerland, June 2018. Laboratoire laussannois d'informatique et statistique textuelle. URL: `https://www.researchgate.net/publication/331907713_Toward_a_Tool_for_Sentiment_Analysis_for_German_Historic_Plays`.

13 Thomas Schmidt, Manuel Burghardt, and Katrin Dennerlein. „Kann man denn auch nicht lachend sehr ernsthaft sein?" – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen. In Georg Vogeler, editor, *Book of Abstracts, DHd 2018*, pages 244–249, Cologne, Germany, 2018. URL: `https://www.researchgate.net/publication/331908052_Kann_man_denn_auch_nicht_lachend_sehr_ernsthaft_sein-Zum_Einsatz_von_Sentiment_Analyse-Verfahren_fur_die_quantitative_Untersuchung_von_Lessings_Dramen`.

14 Thomas Schmidt, Manuel Burghardt, and Katrin Dennerlein. Sentiment annotation of historic german plays: An empirical study on annotation behavior. In Sandra Kübler and Heike Zinsmeister, editors, *Proceedings of the Workshop for Annotation in Digital Humantities (annDH)*, pages 47–52, Sofia, Bulgaria, August 2018. URL: `http://ceur-ws.org/Vol-2155/schmidt.pdf`.

15 Thomas Schmidt, Manuel Burghardt, and Christian Wolff. Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten. In Manuel Burghardt and Claudia Müller-Birn, editors, *INF-DH-2018*, Berlin, Germany, September 2018. Gesellschaft für Informatik e.V. URL: `https://dl.gi.de/handle/20.500.12116/16996`.

16 Thomas Schmidt, Manuel Burghardt, and Christian Wolff. Toward multimodal sentiment analysis of historic plays: A case study with text and audio for lessing's emilia galotti. In *4th Conference of the Association of Digital Humanities in the Nordic Countries (DHN 2019)*, 2019. URL: `https://www.researchgate.net/publication/331907721_Toward_Multimodal_Sentiment_Analysis_of_Historic_Plays_A_Case_Study_with_Text_and_Audio_for_Lessing's_Emilia_Galotti`.

17 Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848, 2010.