Toward a Tool for Sentiment Analysis for German Historic Plays

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

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

With the availability of large amounts of opinionated data through the Internet (social networks, online forums, product reviews, etc.), computational sentiment analysis has become popular in the early 2000s, especially in the context of social media and online reviews (Liu, 2016). Recently sentiment analysis has also found applications in the digital humanities, most notably in the field of literary studies. Sentiment analysis is used for genre classification (Kim et al., 2017), to investigate shifts in the meaning of words (Buechel et al., 2016), to predict the success of novels (Ashok et al., 2013), or to analyse fairy tales (Alm et al., 2005), novels (Kakkonen & Kakkonen, 2011; Jockers, 2015; Jannidis et al., 2016) and drama (Mohammad, 2011; Nalisnick & Baird, 2013). Many of the current projects in this domain use sentiment lexicons. A sentiment lexicon is a list of words with sentiment annotations (positive/negative values). These words are typically referred to as *sentiment bearing words* (SBW). By adding up the number of positive words and subtracting the number of negative words (or polarity annotations on a metric scale), the overall polarity of a text unit can be calculated (Kennedy & Inkpen, 2006).

We present a project on the exploration of different lexicon-based sentiment analysis techniques for the domain of historic, German drama texts, more concretely on a corpus of Lessing's plays. The corpus is composed of twelve plays and was obtained from the *TextGrid*[1] platform. As historic German texts that, at the same time, also use poetic language challenge standard sentiment analysis lexicons, we conducted a systematic evaluation study, to investigate which configuration of dictionaries and NLP tools yields the best results.

We evaluated several combinations of sentiment lexicons and optimization steps:

- Five existing sentiment dictionaries (Remus et al., 2010; Vo et al., 2009; Mohammad & Turney, 2010; Clematide & Klenner, 2010; Waltinger, 2010) for present German, as well as an accumulated combination of all lexicons were evaluated;
- The extension of each of the above lexicons with historical linguistic variants (Jurish, 2012) was evaluated;
- Different types of stopword lists und lists of most frequent words of the corpus (cf. Saif et al., 2014) were evaluated;
- Lemmatization with the pattern lemmatizer (De Smedt & Daelemans, 2012) and the treetagger (Schmid, 1995) was evaluated;

We evaluated the different configurations against a gold standard corpus of 200 single speeches of our corpus. This method of evaluation can be considered rather unique in this branch of sentiment research, as results are typically evaluated by comparing them to well-known observations that are already available from other, oftentimes hermeneutic, scholarly work (cf. Mohammad, 2011; Nalisnick & Baird, 2013).

The gold standard was created in a preliminary annotation study. Five annotators (all fluent in German language) annotated the polarity (positive or negative) of the character speeches. The annotation of the majority of the annotators defines the final polarity of a speech. The measure of agreement between the annotators point to a mediocre agreement (Fleiss' kappa = 0.47; overall agreement in percent = 77%). These results are in line with related studies in the context of narrative

Schmidt, T. & Burghardt, M. (2018). 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.

[1] https://textgridrep.org/ (note: all URLs mentioned in this article were last visited April 13, 2018)

texts (Alm & Sproat, 2005). The final gold standard corpus consists of 139 negative und 61 positive speeches.

We compared the performance of all aforementioned combinations of sentiment and NLP techniques by calculating the overall polarity and by analyzing typical performance metrics such as the accuracy (Gonçalves et al., 2013). During the evaluation study, we found that

- the extension of lexicons with historical linguistic variants and lemmas yields the highest performance boost,
- lexicons with polarity scales (e.g. from -1 to 1) instead of nominal sentiment-annotations (neg/pos) yield consistently better results,
- lexicons that come with explicit lemma and flection forms typically perform better than generic lemmatization tools.

Going through all the metrics, we identified the following combination of techniques as the setup with the best overall performance:

- SentiWS lexicon (Remus et al., 2010),
- no stopword lists,
- pattern lemmatizer,
- extension with historical linguistic variants;

With an overall accuracy of 67%, the performance is above the random baseline, but still considerably worse than in other domains of sentiment analysis (cf. Vinodhini & Chandrasekran, 2012). However, since we use very basic lexicon-based sentiment analysis techniques and the human annotators who produced the gold standard also had severe problems and disagreements concerning the sentiment annotations, we consider these results as promising. We also found that the lower the agreement between annotators for a speech the more likely the sentiment analysis predicts a wrong class. Furthermore, for the gold standard annotation, annotators could only choose between positive and negative; annotations like neutral or mixed were not possible, which aggravates the annotation as well as the automatic prediction. However, other results of our annotation study show that these classes are indeed relevant for our corpus.

To further investigate the possibilities of sentiment analysis in German drama texts, we developed a web application[2] that can be used to explore the results of our current project. Users are able to analyze sentiment progressions and sentiment distributions on several different levels. The structural levels of analysis are the whole drama, single acts, scenes and speeches. Furthermore, by accumulating the speeches of single speakers, users can explore sentiment processes and distributions of specific characters. By using a heuristic described in Nalisnick and Baird (2013), we also integrated sentiment relationships of speakers. Sentiments of speakers and speaker relationships can be analyzed on all structural levels. Besides polarities (positive/negative), we also integrated our results on eight basic emotions as implemented in the NRC Emotion Lexicon (Mohammad & Turney, 2010). To allow for comparisons (e.g. between scenes), users can choose to normalize the results by the number of all words or SBWs.

We are currently working together with literary scholars to further explore requirements for computer-based sentiment analysis in literary studies. We also started a project to acquire more manually annotated data in the context of German historic plays and are also integrating more polarity classes like neutral and mixed in the annotation process. We are planning to use this data for more exact evaluations of the lexicon approach, but also as training data for machine learning approaches to sentiment analysis. Furthermore, we want to extend our current corpus beyond the scope of Lessing's plays, to enable comparisons of authors, genres and periods.

References

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 confer-*

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

ence on empirical methods in natural language processing (pp. 1753-1764).

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.

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.

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/

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).

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). 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.

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).

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.*

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)*