Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques

Jakob Fehle
Media Informatics Group
University of Regensburg
Regensburg, Germany
jakob.fehle@ur.de

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

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

Abstract

We present the results of an evaluation study in the context of lexicon-based sentiment analysis resources for German texts. We have set up a comprehensive compilation of 19 sentiment lexicon resources and 20 sentiment-annotated corpora available for German across multiple domains. In addition to the evaluation of the sentiment lexicons we also investigate the influence of the following preprocessing steps and modifiers: stemming and lemmatization, part-of-speech-tagging, usage of emoticons, stop words removal, usage of valence shifters, intensifiers, and diminishers. We report the best performing lexicons as well as the influence of preprocessing steps and other modifications on average performance across all corpora. We show that larger lexicons with continuous values like *SentiWS* and *SentiMerge* perform best across the domains. The best performing configuration of lexicon and modifications considering the f1-value and accuracy averages across all corpora achieves around 67%. Preprocessing, especially stemming or lemmatization increases the performance consistently on average around 6% and for certain lexicons and configurations up to 16.5% while methods like the usage of valence shifters, intensifiers or diminishers rarely influence overall performance. We discuss domain-specific differences and give recommendations for the selection of lexicons, preprocessing and modifications.

1 Introduction

Sentiment analysis (also often referred to as opinion mining) is a sub-field of *affective computing*, which deals with the detection and analysis of human sentiment and emotions in various application areas like game design (Halbhuber et al., 2019), health (Hartl et al., 2019) and human-computer interaction (Ortloff et al., 2019). Sentiment analysis focuses on text as modality and refers to the task of classifying texts of various lengths concerning polarity (or valence) expressed in the text, meaning whether the sentiment of a text is rather positive or negative (Liu, 2015). Application areas for sentiment analysis in natural language processing (NLP) are social media content (Mäntylä et al., 2018), social sciences (Schmidt et al., 2020b), health (Moßburger et al., 2020), user-generated content (Schmidt et al., 2020a) , digital humanities (Kim and Klinger, 2018a), and human-computer interaction (Schmidt et al., 2020c) to name just a few examples.

Methods for performing sentiment analysis can be divided into two major branches: lexicon-based (also often referred to as rule-based or dictionary-based methods; Taboada et al., 2011) and machine learning (ML)-based approaches. Lexicon-based sentiment analysis uses lexicons consisting of words that are pre-annotated concerning their sentiment expression, which we refer to as sentiment bearing words (SBWs). There are multiple ways to create and acquire such lexicons like crowdsourcing, expert annotations or semi-automatic approaches (cf. Ribeiro et al., 2016). Values of SBWs can either be binary, e.g. +1 (positive) and -1 (negative) (Waltinger, 2010; Mohammad and Turney, 2013) or continuous (e.g. between -3 and + 3) (Remus et al., 2010; Vo et al., 2009; Emerson and Declerck, 2014) to represent differences in sentiment expression across words more precisely. A text can be assigned with an overall polarity by summing up the values of the positively assigned words and subtracting the values of the negative ones. A negative end result points towards a negative and a positive result towards a positive sentiment; a value of 0 is interpreted as neutral (Taboada et al., 2011).

However, developments in ML in the last decade and especially in recent years have led to a dominance of ML-based methods for most NLP-tasks. Current state-of-the-art sentiment analysis regards

sentiment analysis oftentimes as text sequence classification task with three classes (neutral, positive, negative). Current approaches are based on large transformer-based models like BERT and achieve accuracies up to 95% in standardized evaluation settings for English (Nazir et al., 2020; Jindal and Aron, 2021; Dang et al., 2020; González-Carvajal and Garrido-Merchán, 2021) and around 80-90% in German (Wojatzki et al., 2017; Struß et al., 2019; Chan et al., 2020). ML-based methods are dependant of sentiment-annotated corpora and especially for English, an increasing number of sentiment-annotated data-sets that can be used to train algorithms can be found for various domains (Ribeiro et al., 2016; Balazs and Velásquez, 2016; Singh et al., 2020). When compared to each other, modern ML-based methods usually outperform lexicon-based methods, which more recently only serve as baseline for performance comparisons (Dhaoui et al., 2017; Kim and Klinger, 2018b; Khoo and Johnkhan, 2018; Khan et al., 2017; Kharde et al., 2016). Nevertheless, many languages and also special domains lack large annotated corpora necessary for state-of-the art ML-based sentiment analysis. Since lexicon-based methods are not bound to quality and quantity of training data, they are still a common approach for languages (Mukhtar et al., 2018; Al-Ayyoub et al., 2019) and areas (Aung and Myo, 2017) with fewer resources. Furthermore, lexicon-based methods are fast to apply and easy to comprehend which has also led to their popularity in research areas like digital humanities (Kim and Klinger, 2018a; Schmidt et al., 2018b) and especially the sub-field of computational literary studies (Alm and Sproat, 2005; Reagan et al., 2016; Schmidt and Burghardt, 2018a,b; Schmidt, 2019; Schmidt et al., 2019b,c, 2021). For the English language, various research exists evaluating the performance of sentiment lexicons and modifications on multiple corpora (Khan et al., 2017; Ribeiro et al., 2016) or evaluating and surveying lexicons in a context of larger studies including ML-methods (Tsytsarau and Palpanas, 2012; Medhat et al., 2014; Kharde et al., 2016; Singh et al., 2020). Thus, researchers can build upon recommendations and best practices based on this research when selecting sentiment lexicons, preprocessing steps and other modifications. However, to the best of our knowledge, there are no similar resources that provide an exhaustive and systematic listing and evaluation of lexicon-based methods across various sentiment-annotated corpora for the German language. In the following paper we want to address this gap and systematically evaluate lexicon-based techniques for sentiment analysis for German to provide recommendations for the selection of lexicons, preprocessing steps and further configurations. The contributions of this paper are as follows: (1) a comprehensive listing of datasets of sentiment lexicons and sentiment-annotated corpora in German, (2) an in-depth evaluation of resources and methods of lexicon-based sentiment analysis for German, and (3) a discussion of validated recommendations concerning the selection of sentiment lexicons, preprocessing steps and other modifications.

2 Resources

To acquire an exhaustive list of relevant corpora and lexicons for German sentiment analysis we searched in various digital libraries and search engines with appropriate search terms. The most important platforms we investigated are the ACM Digital Library[1], ACL Anthology[2], IEEE[3], Springer Verlag[4] and, on the other hand, more specific platforms such as the Conference on Natural Language Processing[5] (KONVENS). Other sources we referred to are the publications related to the regularly held GermEval[6] competitions or publications of the Interest Group on German Sentiment Analysis[7] (IGGSA). Please note that we do not include resources in the context of German-based emotion analysis. While this research area certainly neighbours sentiment analysis, it is out of scope of this paper. Before discussing the different preprocessing and modification steps, we present an overview of corpora as well as lexicons that we have found for German sentinment analysis.

2.1 Corpora

First, we present all German sentiment annotated corpora we managed to find and that were publicly available or accessible per request (see Table 1). The corpora are of varying quantity and quality. Major differences concern, among other things, the

[1] https://dl.acm.org/
[2] https://www.aclweb.org/anthology/
[3] https://www.ieee.org/
[4] https://www.springer.com/de
[5] https://konvens.org/site/
[6] https://germeval.github.io/
[7] https://sites.google.com/site/iggsahome/

Abbrevation	Corpus name (if reported)	Reference	#Pos	#Neg
LT01-Zehe	German Novel Dataset	Zehe et al., 2017	75	89
LT02-Schmidt		Schmidt et al., 2019a	202	370
LT03-Schmidt		Schmidt et al., 2018a	61	139
MI01-Clematide	MLSA	Clematide et al., 2012	69	110
MI02-Wojatzki	GermEval 2017	Wojatzki et al., 2017	1,537	6,887
MI03-Rauh		Rauh, 2018	333	475
NA01-Butow	GerSEN	Bütow et al., 2016	372	485
NA02-Ploch	GerOM	Ploch, 2015	71	38
NA03-Schabus	One Million Posts Corpus	Schabus et al., 2017	43	1,606
RE01-Klinger	USAGE	Klinger and Cimiano, 2014	506	50
RE02-Sänger	SCARE	Sänger et al., 2016	418,9 k	185,7 k
RE03-Du	SentiLitKrit	Du and Mellmann, 2019	718	290
RE04-Guhr		Guhr et al., 2020	39.6 k	15.4 k
RE05-Prettenhofer		Prettenhofer and Stein, 2010	159,3 k	136,8 k
SM01-Cieliebak	SB10k	Cieliebak et al., 2017	1.717	1.130
SM02-Sidarenka	PotTS	Sidarenka, 2016	3,349	1,510
SM03-Narr		Narr et al., 2012	350	237
SM04-Mozetič		Mozetič et al., 2016	16,5 k	11,7 k
SM05-Siegel	German Irony Corpus	Siegel et al., 2017	49	107
SM06-Momtazi		Momtazi, 2012	278	191

Table 1: Listing of all corpora included in the evaluation. Pos and Neg mark the number of respective annotated text units, acronyms are explained in the text. More information can be found in the appendix (Table 4).

size of the corpora, the granularity of the annotated polarity, the text domain, and also the quality of the annotations. The corpora were classified into five different domains based on the text units they contained: literary and historical texts, texts from or related to news articles, product reviews, social media, and mixed corpora with text units from different domains. For more details about the corpora please refer to Table 4 in the appendix or the specific papers of the corpora. The corpora are further referenced with abbreviations, which are composed of a domain assignment and the primary author of the respective publication (see Table 1). We include three corpora containing literary texts (LT01-LT03), three with mixed types (MI01-MI03), three containing news articles (NA01-NA03), five reviews (RE01-RE05) and six social media content (SM01-SM06). Some of the most well-known corpora of our list are SB10k (SM01-Cieliebak), PotTS (SM02-Sidarenka), USAGE (RE01-Klinger), and the GermEval 2017 corpus (MI02-Wojatzki).

2.2 Lexicons

Table 2 illustrates all lexicons we gathered for this evaluation study. For more details concerning the lexicons please refer to the appendix (Table 5). Please note that some of the lexicons share common word entries or are based in part on other resources. The lexicons are referenced with abbreviations, which are composed of a numeration and the primary author of the respective publication since many lexicons have no explicit names given by the authors. The order of numbers has no specific meaning. There are different versions for some lexicons: 05-Siegel-p and 06-Siegel-m, which focus on words from the *Pressrelations* (Scholz et al., 2012) and *MLSA* (Clematide et al., 2012) datasets, and 08-Takamura-c and 09-Takamura-d, respectively, for continuous and dichotomous sentiment values. Several well-known and often used lexicons are also included, such as *SentiWS* (01-Remus), *BAWL-R* (03-Võ), *GermanPolarityClues* (13-Waltinger), and *LIWC-De* (14-Wolf). Our general calculation of sentiment values is as follows: For a text unit, we count the positive and negative matches and subtract the sum of positive words by the negative ones. A positive end result is counted as positive polarity, a negative as a negative one. Across chapter 3 we detail some further methods to adjust this calculation.

Abbrevation	Lexicon name (if reported)	Reference	Tokens
01-Remus	SentiWS	Remus et al., 2010	34,238
02-Clematide		Clematide et al., 2010	9,239
03-Võ	BAWL-R	Vo et al., 2009	2,902
04-Emerson	SentiMerge	Emerson and Declerck, 2014	96,420
05-Siegel-p		Siegel and Diwisch, 2014	2,917
06-Siegel-m		Siegel and Diwisch, 2014	2,917
07-Rill	SePL	Rill et al., 2012	14,395
08-Takamura-c	GermanSentiSpin	Takamura et al., 2005	105,560
09-Takamura-d	GermanSentiSpin	Takamura et al., 2005	88,925
10-Rauh		Rauh, 2018	37,080
11-Du	SentiLitKrit	Du and Mellmann, 2019	3,620
12-Asgari	UniSent	Asgari et al., 2019	1,384
13-Waltinger	GermanPolarityClues	Waltinger, 2010	38,901
14-Wolf	LIWC-De	Wolf et al., 2008	4,894
15-Klinger	USAGE Sentiment Lexicon	Klinger and Cimiano, 2014	4,743
16-Wilson	GermanSubjectivityClues	Wilson et al., 2009	9,827
17-Mohammad	NRC Emotion Lexicon	Mohammad and Turney, 2013	10,617
18-Ruppenhofer		Ruppenhofer et al., 2017	9,544
19-Chen	Multilingual Sentiment Lexicon	Chen and Skiena, 2014	3,973

Table 2: Listing of all lexicons included in our evaluation. Lexicons 1-8 include two versions: dichotomous and continuous sentiment values. The rest is solely dichotomous. More information can be found in the appendix (Table 5).

3 Methods

3.1 General Data Cleaning

We perform the following steps to clean the texts of all corpora before evaluation:

- Removing non-alphabetic characters (numbers, special characters, etc.) as well as leading, trailing and multiple spaces (Haddi et al., 2013).
- The removal of URL links, Twitter usernames, and Twitter-specific words such as "RT" (Pak and Paroubek, 2010).

All of the above steps showed no relevant influence on SBWs or lexicon-based sentiment analysis and serve only normalization purposes.

3.2 Preprocessing and other Modifications

In addition to the evaluation of lexicon resources, we also investigate the influence on performance by various preprocessing steps and other configurations which are frequently used when preparing the application of sentiment lexicons. The following techniques are evaluated: The assignment and use of part-of-speech (POS) information, lemmatization and stemming, emoticon processing, stop words removal, lowercasing and the application of valence-changing words. We will refer to these techniques in the following as *modifiers* or *modifications*. Most modifiers are either on or off, meaning they are performed or not, except for POS-tagging, stemming and lemmatization for which multiple approaches are evaluated as well as on and off. In order to identify the best combination of modifiers in the context of the chosen lexicon, the different methods are cross-evaluated and compared based on classification metrics.

3.2.1 Part-of-Speech-Tagging

In sentiment analysis, POS information can be used to solve the problem of word ambiguity since words with the same spelling can have a different valence dependent of the POS (Taboada et al., 2011). Knowledge of the correct POS can support the resolving of this kind of ambiguity. It is necessary to perform POS-tagging on the text and on the lexicon (few of our lexicons already do contain POS information). We evaluate and use two of the most well-known POS-taggers for German: *TreeTagger* (Schmid, 2013) which has shown good performance in evaluation studies (Gleim et al., 2019; Horsmann et al., 2015) and *Stanza* (Qi et al., 2020), a novel POS-tagger for German. Sentiment

lexicons consist almost exclusively of nouns, adjectives, verbs and adverbs, which are mainly responsible for the polarity of a text unit (Pak and Paroubek, 2010). Therefore, all POS information was normalized to these four categories. When we apply POS-tagging in our sentiment analysis pipeline, after finding matching words between text and lexicon, we also test if the POS matches or refers to the word with the correct POS before including it in the calculation.

3.2.2 Stemming or Lemmatization

While some sentiment lexicons contain various inflections of words (Remus et al., 2010), the vocabulary of these lexicons mostly consist of base forms. To enable the mapping of words in texts and in the lexicon, base form reduction via lemmatization or stemming is often applied (Taboada et al., 2011). Stemming refers to algorithms that attempt to reduce the word to the base form by truncating suffixes and affixes based on predefined rules. Lemmatization, on the other hand, often takes sentence order and surrounding words into account or works with large dictionaries to reduce a word to its true base form, the lemma, which is necessary for languages with complex morphology like German. In this study, we evaluate the usage of the following two lemmatizers for German: *TreeTagger* (Schmid, 2013), and *Inverse Wiktionary for Natural Language Processing* (IWNLP) (Liebeck and Conrad, 2015). In terms of stemming, two established stemming algorithms are evaluated: *Cistem* (Weissweiler and Fraser, 2017) and *Snowball Porter* (Porter, 1980). Please note that we do not evaluate the lemmatizers or stemming approaches for their intended task but only with respect to the influence on sentiment analysis (the same holds true for POS-tagging). For a review of base form reduction in German we recommend Gleim et al. (2019). We evaluate these methods by applying stemming/lemmatization to the text and lexicon before looking for the matches.

3.2.3 Lowercasing

Unlike English, German does not only capitalize the beginning of sentences and proper names, but also nouns or nominalizations. Thus, for certain cases, it is important to differ between cased and uncased versions of words in German to disambiguate sentiment (e.g. "würde" (*would*, auxiliary verb) has no sentiment, "Würde" (*dignity*, noun) is positive in some lexicons). However, written text in general and in social media in particular includes a lot of spelling errors and incorrect capitalization hindering correct sentiment calculations. Therefore, we evaluate how lowercasing of the lexicon and the texts influences performance.

3.2.4 Emoticons

Emoticons are representations of body language in text, very frequently connected to sentiment expressions (Ptaszynski et al., 2011). Since emoticons are common on the social web, several papers show the benefits of including emoticons in the calculation of the sentiment value of a text unit (Hogenboom et al., 2015; Gonçalves et al., 2013). To translate emoticons to sentiment values, we used a 232-entry list of emoticons from the SCARE dataset by Sänger et al. (2016). Positive or negative emoticons are treated as additional entries to the lexicon vocabulary (positive as +1, negative as -1).

3.2.5 Stop Words Removal

The removal of stop words, i.e. common words (like function words) that occur with high frequency in a language, is a common practice in NLP pipelines, predominantly to improve computation performance. In this process, the individual words of a text unit are matched against a list of words and removed from the text unit if they match any of the entries. Common stop words in language are articles, prepositions, conjunctions, and pronouns and they usually bear no sentiment. While stop words usually have no influence on calculations via lexicon-based methods, sentiment lexicons that are created automatically or semi-automatically can contain stop words which can skew sentiment calculations, e.g. "dieser, jetzt, ihnen, ihrer, ihm" in the lexicon 08-Takamura-d. Such entries are not considered further by removing stop words. Indeed, in some settings the removal of stop words has been shown to be beneficial for sentiment analysis (Saif et al., 2014). We evaluate the application of the German stop words list provided by the information retrieval framework *Solr*.[8]. The list is rather conservative with a length of 231 entries. If we use the modification stop words list, words of this list are ignored in the text as well as in the lexicon that is used.

3.2.6 Valence Shifters

Depending on the surrounding of a SBW, the sentiment value of a word can be influenced, for exam-

[8] https://solr.apache.org/

ple the word "glücklich" (happy), usually positive, turns negative with the negation "nicht" (not) right before. Such words and phrases are referred to as valence shifters (Mohammad, 2016). It is recommended to include valence shifters into the calculation process for lexicon-based sentiment analysis (Pröllochs et al., 2015). The following parameters are important for dealing with valence shifters: (1) the window size, meaning how close a valence shifter has to be to a SBW to influence calculations and (2) the position, meaning if the valence shifter is left or right of the SBW (Pang et al., 2002; Kennedy and Inkpen, 2006). For this work, we used a two-sided window with a fixed length of 4 words, which achieved the best results in a wider comparison of methods on German-language datasets by Pröllochs et al. (2015). If a valence shifter occurs in the text, the sentiment values of all words within the context window are reversed. We use a list of 22 German negations collected by various lists (Clematide et al., 2010; Ruppenhofer et al., 2017; Tymann et al., 2019).

3.2.7 Valence Intensifiers and Diminishers

Similar to valence shifters, words can also act as valence intensifiers or diminishers e.g. "sehr" (very) or "wenig" (little). As with valence shifters, a variety of possible implementation approaches exist regarding the context window and position of these words (Taboada et al., 2011; Klenner et al., 2009. We chose to use the approach of Taboada et al. (2011): given a context window of 2 words before the SBW, the sentiment values of all SBWs within the window are multiplied by the value of the diminishers or intensifier. We use a list of 78 German intensifiers and diminishers by Clematide et al. (2010) and Ruppenhofer et al. (2017).

3.2.8 Usage of Lexicon-specific continuous Sentiment Values

While most lexicons have sentiment values in dichotomous (positive, negative) or trichotomous (positive, negative, neutral) expressions, some lexicons contain sentiment values with continuous values, for example between +3 and -3. Thus, if a lexicon offers continuous metrics, we evaluate both approaches: the usage of these continuous values in the calculation and the binary representation via +1 and -1. This is the case for the lexicons 1-8.

4 Results

We evaluate the lexicons and modifiers regarding sentiment analysis as binary classification tasks with positive and negative values, ignoring all neutral information. If a calculation produces 0 (neutral) as output, this is counted as false prediction. In chapter 4.1 we first present the lexicon performance without using modifiers to investigate the general performance of lexicons, corpora and domains. In chapter 4.2 we present modifier-based results before we take a closer look at the best lexicon-modifier combinations in chapter 4.3.

Due to the high class imbalances of certain corpora, we primarily report macro f1 measure. When we report averages across corpora we do not account for size imbalances of the corpora. Instead we calculate the mean average of f1 measures over all corpora.[9]

4.1 Lexicon Performance without modifiers

First, we present the results of cross-evaluations when using the sentiment lexicons on the corpora without any modifiers via a heatmap (see Fig. 1). Please note that the random and majority baselines of the corpora fluctuate around 50-70% for most corpora (see 4 in the appendix). The average f1 measure of all lexicons across all corpora is 45%. A few lexicons achieve an average f1 measure above 50% across all corpora. The best performing lexicons are, on average, 13-Waltinger with 60%, 10-Rauh with 57%, 19-Chen with 53%, 01-Remus with 52% and 04-Emerson with 51%. However, multiple lexicons do perform way below 50%. Considering differences on the corpora of various domains, we have identified the following findings: On average, the lexicons perform best on corpora from the product review domain with f1 scores between 46 to 58%. Corpora based on social media content lead to rather low f1 values between 36-46%. The f1 scores do however vary a lot, for certain lexicons around 10-20% showing that the selection of the corpus-lexicon combination is important. The best result is achieved by lexicons designed specifically for the task on certain corpora e.g. 15-Klinger on corpus RE01-Klinger with

[9] We limit the result report to the most important results. However, we publish a GitHub repository including all results for all lexicon-modifier combination across all corpora for multiple performance metrics and further overview data like heat maps and domain specific result tables. The repository can be found at https://github.com/JakobFehle/Lexicon-based-SentA-German

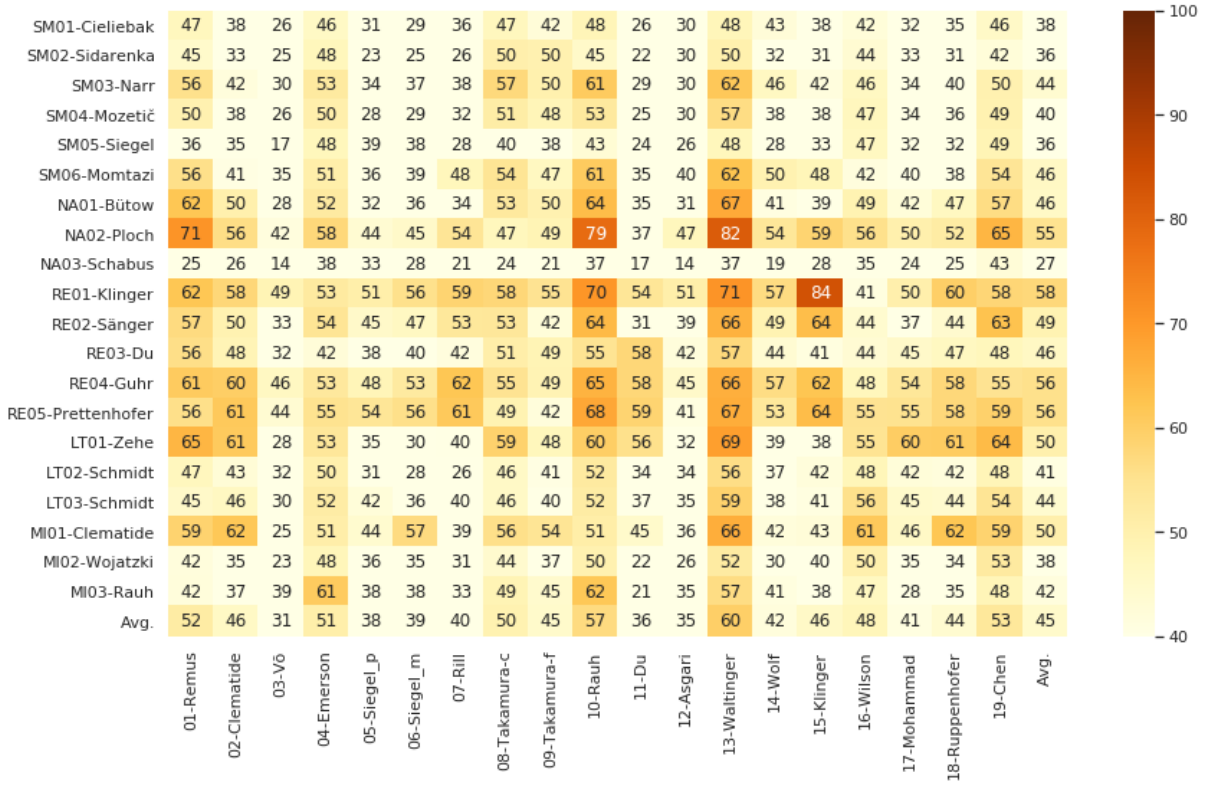

Figure 1: Heatmap for the cross-evaluation of lexicons and corpora including overall averages with no modifications. Values are given as f1 measure and rounded. X-axis are the lexicons, y-axis the corpora.

84%. Other good performances are found with 13-Waltinger and 10-Rauh on NA02-Ploch with 82% and 79%, respectively.

4.2 Modifications

To evaluate the effects of the respective modifiers, they are examined in two ways: (1) We regard the average performance of all lexicons and corpora without modifiers as baseline (f1 measure of 44.8%) and compare it with the average of the isolated use of a single modifier turned on (all other modifiers off) across all corpora and lexicons. We refer to the difference of the baseline f1 measure (44.8%) and the average across all corpora and lexicons with just this modifier turned on as *f1-raw-delta*. A positive value shows an improvement, a negative value a decrease of performance. (2) We measure every possible on/off configuration for all modifiers across all corpora and lexicons once with the specific examined modifier on and once off. We then take the multiple differences between modifier on and off for all of this runs and build an average. We refer to this value as *f1-combination-delta*. Please refer to Table 6 in the appendix for a detailed overview of the results.

Concerning POS-tagging, stemming and lemmatization, the different tools show very low differences. Therefore, we always refer to the best-performing tool as representative of the method. POS-tagging leads to a small decrease of the f1 measure compared to not applying it (f1-raw-delta = -1.7%) and also on average combined with other modifiers (f1-combination-delta = -1.5%). Stemming and lemmatization however improves f1 measures and is the most consistent and strongest improvement. F1-raw-delta shows an improvement by 6.3% for the best stemming-method and 5.6% for the best lemmatizer. This result stays consistent for f1-combination-delta with 5% and 5.1% respectively.

Lowercasing shows a smaller positive influence (f1-raw-delta = 2.8; f1-combination-delta = 0.2). Including emoticons in the calculation process improves the performance similarly but also consistent in combination with other modifiers (f1-raw-delta = 2.9; f1-combination-delta = 1.7). The increase of the f1 measure is connected to the corpora of the social media domain. The processing of emoticons improves the f1 measure actually by 8.8% points when we reduce the results on the social media corpora. The removal of stop words be-

fore performing calculation does actually decrease the average f1 measure by 0.3% when no other methods are applied. Intertwined with other methods, this decrease is also marginal (f1-combination-delta = -0.2). Integrating valence shifters into calculation does actually barely show an influence on performance according to our evaluations (f1-raw-delta = 0.0; f1-combination-delta = -0.2). The same holds true for intensifiers and diminushers (f1-raw-delta = 0.5; f1-combination-delta = 0.4).

For the modifier of continuous sentiment values, we limit the calculation of f1-raw-delta to the 8 lexicons containing such values, thus the baseline is 43.3%. The application of continuous values improves the f1-score by 3.4%. Indeed an improvement can be found for every lexicon compared to their dichotomous equivalent.

Please note however that the values given above are averaged overall results. Several methods do actually have much higher positive influence depending on the specific corpus-lexicon combination. The following sub-chapter will highlight some of these interaction effects.

4.3 Lexicon Performance with Modifications

In the following chapter, we present the best result achieved with various modifier combinations for each lexicon (see Table 3). Next to the highest f1 measure, we also report the average performance (averaging the result of all method combinations). For the lexicons 1-8 we differ between the continuous and the dichotomous calculation (the latter in brackets in Table 3). More information about the precise combination of methods can be found in the appendix in Table 7.

Lexicon 04-Emerson achieves both the highest average performance with an f1 measure of 62.0% over all method combinations and the best specific combination with 67.3% in regards to all lexicons and all method combinations. The modifiers are lemmatization (IWNLP lemmatizer), lowercasing, removing stop words, using emoticons as well as continuous sentiment values; all other modifiers are turned off. This value is 16.5% higher than the baseline of 04-Emerson using no modifier showing that in contrast to the overall results in chapter 4.1, certain modifier combinations can highly boost performance.

The f1 values for all lexicons range between 52 and 67% for the best methods. Overall, the best performing lexicons with no modifications are mostly

Lexicon Performance		
Lexicon	Average-f1	Best-Method-f1
04-Emerson	*62.0* (55.1)	67.3
01-Remus	*60.1* (55.1)	63.6
10-Rauh	58.5	63.6
02-Clematide	*56.5* (54.6)	61.9
08-Takamura-c	*56.3* (52.3)	60.6
19-Chen	55.8	60.5
13-Waltinger	55.2	63.4
18-Ruppenhofer	53.5	59.2
16-Wilson	52.2	56.0
07-Rill	*51.4* (48.7)	56.2
03-Võ	*49.3* (45.0)	54.9
09-Takamura-d	48.8	52.7
14-Wolf	48.6	53.9
06-Siegel_m	*47.7* (45.9)	53.8
17-Mohammad	47.4	51.6
15-Klinger	46.8	53.4
05-Siegel_p	*46.8* (45.9)	54.3
11-Du	46.7	53.2
12-Asgari	43.9	52.0

Table 3: Lexicon performance in combination with modifiers. Best method is the value for the best modifier combination for each lexicon. Average is the overall average for all modifier combinations of this lexicon. Values in brackets are results for dichotomous equivalents for lexicons 1-8.

the same as with modifications (see chapter 4.1, 4.3 and Fig. 1) but modifiers increase the performance by 5-17%. The best combination for each lexicon consistently includes emoticons and stemming or lemmatizing. Four of the best five performing lexicons work with continuous values. Considering lowercasing, stop words removal, valence shifters, intensifier and diminushers, the usage is rather inconsistent among the best lexicon-modifier combinations. POS-tags are only part of the best combination for 10-Rauh (see Table 7 in the appendix).

5 Discussion

In the following chapter, we summarize the results and formulate recommendations and best practices for the usage of German general purpose sentiment lexicons. We have evaluated, to our knowledge, all relevant and publicly available corpora and lexicons for the German language. The six best performing lexicons without preprocessing and modifications but also with such methods are: *SentiMerge* (*04-Emerson*) (Emerson and Declerck, 2014), *Sen-*

tiWS (*01-Remus*) (Remus et al., 2010), 10-Rauh (Rauh, 2018), the *Multilinguial Sentiment Lexicon* (19-Chen) (Chen and Skiena, 2014), 02-Clematide (Clematide et al., 2010) and *GermanSentiSpin* (08-Takamura-c) (Takamura et al., 2005). Performance can vary a lot depending of domain and corpus, however these lexicons perform, on average, well on all domains compared to the other evaluated lexicons. Therefore, we recommend the usage of these lexicons. *SentiMerge* (04-Emerson) achieves the best result with a specific modifier setting (f1 measure = 67.3%), thus we especially encourage the usage of this lexicon. On average, larger lexicons (that consist of more entries) perform better. Indeed, 04-Emerson is the second largest resource in our evaluation, although there are exceptions. Lexicons performing rather good but which are small: e.g. 02-Clematide and 19-Chen. It is striking that 04-Emerson is actually a lexicon derived by fusing multiple other lexicons to increase items size (Emerson and Declerck, 2014). We recommend exploring this idea further in future work. Another pattern that emerges is that on average lexicons with continuous sentiment values outperform dichotomous annotations, which has also been shown in other studies for English (Taboada et al., 2011). Based on these result we conclude that continuous representations of sentiment expressions fit human language more.

Considering modifications and preprocessing, we have identified that the application of one single modifier rarely helps, and we recommend the combination of multiple modifications and preprocessing steps. The most consistent and supportive modifier is the application of stemming or lemmatization of lexicon and text which solves the problem of complex inflections matching in the sentiment analysis pipeline. We did not identify a large difference between these two methods or between specific tools implementing them. POS-tagging, on the other hand showed no significant improvement.

Another consistent boost is the integration of emoticons into the calculation, especially for tasks in the social media area (Hogenboom et al., 2013; Pozzi et al., 2013). The removal of stop words and lowercasing produced inconsistent results. Overall, the modifications are not necessary or beneficial based on our results. In contrast to previous research on German (Pröllochs et al., 2015), we could not identify an improvement by integrating valence shifters, intensifiers and diminishers into our calculation. This result is counter-intuitive; we assume that the specific selection of a larger window size and position (see chapter 3.2.6) might be a reason for this. We plan to investigate this phenomenon in future work in more detail, but cannot recommend the application of these modifiers the way we did in this evaluation.

With regard to corpora and domains, we identified that, as expected, lexicons that are designed for specific corpora or domains perform best on these corpora. Overall, the evaluated lexicons perform best on product reviews while social media corpora are more challenging. We encourage to address these problems in future work in sentiment analysis.

Summing up, we must note that compared to English lexicon-based resources which can achieve f1 measures above 70% (Khan et al., 2017; Ribeiro et al., 2016) the German resources perform rather poorly. German resources often lack size and suffer from strong class imbalances resulting in the sometimes fairly poor results reported here. This accounts for lexicons as well as for corpora and influences performance negatively. The rise of ML-based methods and their better performance compared to lexicon-based methods will certainly hinder the further development and improvement of sentiment lexicons. However, as the popularity of resources like VADER (Hutto and Gilbert, 2014) for English language shows, there is still an interest by certain communities for fast and easy-to-use sentiment lexicons to perform sentiment analysis. Thus, we not just want to support decision-making for German resources with the presented evaluation study, but give impulses for future developments for German sentiment analysis resources.

References

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. *Information processing & management*, 56(2):320–342.

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional Sequencing and Development in Fairy Tales. In *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science, pages 668–674, Berlin, Heidelberg. Springer.

Ehsaneddin Asgari, Fabienne Braune, Benjamin Roth, Christoph Ringlstetter, and Mohammad RK Mofrad. 2019. Unisent: Universal adaptable sentiment

lexica for 1000+ languages. *arXiv preprint arXiv:1904.09678*.

Khin Zezawar Aung and Nyein Nyein Myo. 2017. Sentiment analysis of students' comment using lexicon based approach. In *2017 IEEE/ACIS 16th international conference on computer and information science (ICIS)*, pages 149–154. IEEE.

Jorge A Balazs and Juan D Velásquez. 2016. Opinion mining and information fusion: a survey. *Information Fusion*, 27:95–110.

Florian Bütow, Andreas Lommatzsch, and Danuta Ploch. 2016. Creation of a german corpus for internet news sentiment analysis.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. *arXiv:2010.10906 [cs]*. ArXiv: 2010.10906.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. Mlsa—a multi-layered reference corpus for german sentiment analysis.

Simon Clematide, Manfred Klenner, A Montoyo, P Martínez-Barco, A Balahur, and E Boldrini. 2010. Evaluation and extension of a polarity lexicon for german.

Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

Chedia Dhaoui, Cynthia M Webster, and Lay Peng Tan. 2017. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*.

Keli Du and Katja Mellmann. 2019. Sentimentanalyse als instrument literaturgeschichtlicher rezeptionsforschung. ein pilotprojekt.

Guy Emerson and Thierry Declerck. 2014. Sentimerge: Combining sentiment lexicons in a bayesian framework. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 30–38.

Rüdiger Gleim, Steffen Eger, Alexander Mehler, Tolga Uslu, Wahed Hemati, Andy Lücking, Alexander Henlein, Sven Kahlsdorf, and Armin Hoenen. 2019. A practitioner's view: a survey and comparison of lemmatization and morphological tagging in german and latin. *Journal of Language Modelling*, 7.

Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. 2013. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38.

Santiago González-Carvajal and Eduardo C. Garrido-Merchán. 2021. Comparing bert against traditional machine learning text classification.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1627–1632.

Emma Haddi, Xiaohui Liu, and Yong Shi. 2013. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26–32.

David Halbhuber, Jakob Fehle, Alexander Kalus, Konstantin Seitz, Martin Kocur, Thomas Schmidt, and Christian Wolff. 2019. The mood game - how to use the player's affective state in a shoot'em up avoiding frustration and boredom. In *Proceedings of Mensch Und Computer 2019*, MuC'19, page 867–870, New York, NY, USA. Association for Computing Machinery.

Philipp Hartl, Thomas Fischer, Andreas Hilzenthaler, Martin Kocur, and Thomas Schmidt. 2019. Audiencear - utilising augmented reality and emotion tracking to address fear of speech. In *Proceedings of Mensch Und Computer 2019*, MuC'19, page 913–916, New York, NY, USA. Association for Computing Machinery.

Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1&2):22–40.

Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. 2013. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th annual ACM symposium on applied computing*, pages 703–710.

Tobias Horsmann, Nicolai Erbs, and Torsten Zesch. 2015. Fast or accurate?-a comparative evaluation of pos tagging models. In *GSCL*, pages 22–30.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Kanika Jindal and Rajni Aron. 2021. A systematic study of sentiment analysis for social media data. *Materials Today: Proceedings*.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Farhan Hassan Khan, Usman Qamar, and Saba Bashir. 2017. Lexicon based semantic detection of sentiments using expected likelihood estimate smoothed odds ratio. *Artificial Intelligence Review*, 48(1):113–138.

Vishal Kharde, Prof Sonawane, et al. 2016. Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.

Christopher SG Khoo and Sathik Basha Johnkhan. 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Evgeny Kim and Roman Klinger. 2018a. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.

Evgeny Kim and Roman Klinger. 2018b. 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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. Polart: A robust tool for sentiment analysis. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 235–238.

Roman Klinger and Philipp Cimiano. 2014. The usage review corpus for fine-grained, multi-lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Matthias Liebeck and Stefan Conrad. 2015. Iwnlp: Inverse wiktionary for natural language processing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 414–418.

Bing Liu. 2015. Sentiment analysis: mining opinions, sentiments, and emotions.

Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Saeedeh Momtazi. 2012. Fine-grained german sentiment analysis on social media. In *LREC*, pages 1215–1220. Citeseer.

Luis Moßburger, Felix Wende, Kay Brinkmann, and Thomas Schmidt. 2020. Exploring online depression forums via text mining: A comparison of Reddit and a curated online forum. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 70–81, Barcelona, Spain (Online). Association for Computational Linguistics.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Neelam Mukhtar, Mohammad Abid Khan, and Nadia Chiragh. 2018. Lexicon-based approach outperforms supervised machine learning approach for urdu sentiment analysis in multiple domains. *Telematics and Informatics*, 35(8):2173–2183.

Sascha Narr, Michael Hulfenhaus, and Sahin Albayrak. 2012. Language-independent twitter sentiment analysis. *Knowledge discovery and machine learning (KDML), LWA*, pages 12–14.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*.

Anna-Marie Ortloff, Lydia Güntner, Maximiliane Windl, Thomas Schmidt, Martin Kocur, and Christian Wolff. 2019. Sentibooks: Enhancing audiobooks via affective computing and smart light bulbs. In *Proceedings of Mensch Und Computer 2019*, MuC'19, page 863–866, New York, NY, USA. Association for Computing Machinery.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Danuta Ploch. 2015. Intelligent news aggregator for german with sentiment analysis. In *Smart information systems*, pages 5–46. Springer.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*.

Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Daniele Blanc. 2013. Enhance polarity classification on social media through sentiment-based feature expansion. *WOA@ AI* IA*, 1099:78–84.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127.

Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. 2015. Enhancing sentiment analysis of financial news by detecting negation scopes. In *2015 48th Hawaii International Conference on System Sciences*, pages 959–968. IEEE.

Michal Ptaszynski, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. 2011. Research on emoticons: review of the field and proposal of research framework. *Proceedings of 17th Association for Natural Language Processing*, pages 1159–1162.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Christian Rauh. 2018. Validating a sentiment dictionary for german political language—a workbench note. *Journal of Information Technology & Politics*, 15(4):319–343.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31. ArXiv: 1606.07772.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. Sentiws-a publicly available german-language resource for sentiment analysis. In *LREC*. Citeseer.

Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29.

Sven Rill, Jörg Scheidt, Johannes Drescher, Oliver Schütz, Dirk Reinel, and Florian Wogenstein. 2012. A generic approach to generate opinion lists of phrases for opinion mining applications. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, pages 1–8.

Josef Ruppenhofer, Petra Steiner, and Michael Wiegand. 2017. Evaluating the morphological compositionality of polarity.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 810–817, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mario Sänger, Ulf Leser, Steffen Kemmerer, Peter Adolphs, and Roman Klinger. 2016. Scare—the sentiment corpus of app reviews with fine-grained annotations in german. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1114–1121.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1241–1244.

Helmut Schmid. 2013. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154.

Thomas Schmidt. 2019. Distant reading sentiments and emotions in historic german plays. In *Abstract Booklet, DH_Budapest_2019*, pages 57–60. Budapest, Hungary.

Thomas Schmidt and Manuel Burghardt. 2018a. 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, Santa Fe, New Mexico. Association for Computational Linguistics.

Thomas Schmidt and Manuel Burghardt. 2018b. Toward a Tool for Sentiment Analysis for German Historic Plays. In *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018*, pages 46–48, Lausanne, Switzerland. Laboratoire laussannois d'informatique et statistique textuelle.

Thomas Schmidt, Manuel Burghardt, and Katrin Dennerlein. 2018a. Sentiment annotation of historic german plays: An empirical study on annotation behavior. In Sandra Kübler and Heike Zinsmeister, editors, *annDH 2018, Proceedings of the Workshop on Annotation in Digital Humanities 2018 (annDH 2018), Sofia, Bulgaria, August 6-10, 2018*, pages 47–52. RWTH Aachen, Aachen.

Thomas Schmidt, Manuel Burghardt, Katrin Dennerlein, and Christian Wolff. 2019a. Sentiment annotation for lessing's plays: Towards a language resource for sentiment analysis on german literary

texts. In Thierry Declerck and John P. McCrae, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, pages 45–50. RWTH Aachen, Aachen.

Thomas Schmidt, Manuel Burghardt, and Christian Wolff. 2018b. Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten. In *INF-DH-2018*, Berlin, Germany. Gesellschaft für Informatik e.V.

Thomas Schmidt, Manuel Burghardt, and Christian Wolff. 2019b. Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, volume 2364 of *CEUR Workshop Proceedings*, pages 405–414, Copenhagen, Denmark. CEUR-WS.org.

Thomas Schmidt, Johanna Dangel, and Christian Wolff. 2021. Senttext: A tool for lexicon-based sentiment analysis in digital humanities. In Thomas Schmidt and Christian Wolff, editors, *Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)*, volume 74, pages 156–172. Werner Hülsbusch, Glückstadt.

Thomas Schmidt, Philipp Hartl, Dominik Ramsauer, Thomas Fischer, Andreas Hilzenthaler, and Christian Wolff. 2020a. Acquisition and analysis of a meme corpus to investigate web culture. In *Digital Humanities Conference 2020 (DH 2020*.

Thomas Schmidt, Florian Kaindl, and Christian Wolff. 2020b. Distant reading of religious online communities: A case study for three religious forums on reddit. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pages 157–172, Riga, Latvia.

Thomas Schmidt, Miriam Schlindwein, Katharina Lichtner, and Christian Wolff. 2020c. Investigating the relationship between emotion recognition software and usability metrics. *i-com*, 19(2):139–151.

Thomas Schmidt, Brigitte Winterl, Milena Maul, Alina Schark, Andrea Vlad, and Christian Wolff. 2019c. Inter-rater agreement and usability: A comparative evaluation of annotation tools for sentiment annotation. In *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*, pages 121–133, Bonn. Gesellschaft für Informatik e.V.

Thomas Scholz, Stefan Conrad, and Lutz Hillekamps. 2012. Opinion mining on a german corpus of a media response analysis. In *International Conference on Text, Speech and Dialogue*, pages 39–46. Springer.

Uladzimir Sidarenka. 2016. Potts: the potsdam twitter sentiment corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1133–1141.

Melanie Siegel and Kerstin Diwisch. 2014. Opm. https://sites.google.com/site/iggsahome/downloads. Last access: 05.12.2020.

Melanie Siegel, Katharina Emig, Nikola Ihringer, Serhat Kesim, and Tamer Yilmaz. 2017. Sentiment analysis. ressources for sentiment analysis of german language. https://github.com/hdaSprachtechnologie/Sentiment-Analysis. Last access: 10.12.2020.

Nikhil Kumar Singh, Deepak Singh Tomar, and Arun Kumar Sangaiah. 2020. Sentiment analysis: a review and comparative analysis over social media. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):97–117.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 133–140.

Mikalai Tsytsarau and Themis Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

Karsten Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. 2019. Gervader-a german adaptation of the vader sentiment analysis tool for social media texts. In *LWDA*, pages 178–189.

Melissa LH Vo, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41(2):534–538.

Ulli Waltinger. 2010. Germanpolarityclues: A lexical resource for german sentiment analysis. In *LREC*, pages 1638–1642. Citeseer.

Leonie Weissweiler and Alexander Fraser. 2017. Developing a stemmer for german based on a comparative analysis of publicly available stemmers. In *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 81–94. Springer.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, pages 1–12.

Markus Wolf, Andrea B Horn, Matthias R Mehl, Severin Haug, James W Pennebaker, and Hans Kordy. 2008. Computergestützte quantitative textanalyse: äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica*, 54(2):85–98.

Albin Zehe, Martin Becker, Fotis Jannidis, and Andreas Hotho. 2017. Towards sentiment analysis on german literature. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 387–394. Springer.

A Appendix

Abbreviation	Domain	Corpus	Reference	Pos	Neg	Total
LT01-Zehe	Literary Texts	German Novel Dataset	(Zehe et al., 2017)	75	89	164
LT02-Schmidt	Literary Texts		(Schmidt et al., 2019a)	202	370	572
LT03-Schmidt	Literary Texts		(Schmidt et al., 2018a)	61	139	200
MI01-Clematide	Mixed Domains	MLSA	(Clematide et al., 2012)	69	110	179
MI02-Wojatzki	Mixed Domains	GermEval 2017	(Wojatzki et al., 2017)	1,537	6,887	8,424
MI03-Rauh	Mixed Domains		(Rauh, 2018)	333	475	808
NA01-Bütow	News Articles	GerSEN	(Bütow et al., 2016)	372	485	857
NA02-Ploch	News Articles	GerOM	(Ploch, 2015)	71	38	109
NA03-Schabus	News Articles	One Million Posts Corpus	(Schabus et al., 2017)	43	1,606	1,649
RE01-Klinger	Product Reviews	USAGE	(Klinger and Cimiano, 2014)	506	50	556
RE02-Sänger	Product Reviews	SCARE	(Sänger et al., 2016)	418,880	185,666	604,546
RE03-Du	Product Reviews	SentiLitKrit	(Du and Mellmann, 2019)	718	290	1,008
RE04-Guhr	Product Reviews		(Guhr et al., 2020)	39,623	15,436	55,059
RE05-Prettenhofer	Product Reviews		(Prettenhofer and Stein, 2010)	159,315	136,757	296,072
SM01-Cieliebak	Social Media	SB10k	(Cieliebak et al., 2017)	1,717	1,130	2,847
SM02-Sidarenka	Social Media	PotTS	(Sidarenka, 2016)	3,349	1,510	4,859
SM03-Narr	Social Media		(Narr et al., 2012)	350	237	587
SM04-Mozetič	Social Media		(Mozetič et al., 2016)	16,502	11,693	28,195
SM05-Siegel	Social Media	German Irony Corpus	(Siegel et al., 2017)	49	107	156
SM06-Momtazi	Social Media		(Momtazi, 2012)	278	191	469

Table 4: List of corpora with information about the respective domain of the texts, information about the size of the corpora, an abbreviation for further identification, and optionally, a corpus name if given by the authors. Pos and Neg illustrate the number of text units for the respective sentiment class in the corpus.

Abbreviation	Lexicon	Reference	Pos. Tokens	Neu. Token	Neg. Tokens	Total Tokens	Cont. Val.
01-Remus	SentiWS	(Remus et al., 2010)	16,385	0	17,853	34,238	x
02-Clematide		(Clematide et al., 2010)	3,378	608	5,253	9,239	x
03-Võ	BAWL-R	(Vo et al., 2009)	1,576	60	1,266	2,902	x
04-Emerson	SentiMerge	(Emerson and Declerck, 2014)	45,301	221	50,898	96,420	x
05-Siegel-p		(Siegel and Diwisch, 2014)	1,142	365	1,410	2,917	x
06-Siegel-m		(Siegel and Diwisch, 2014)	1,104	396	1,417	2,917	x
07-Rill	SePL	(Rill et al., 2012)	11,015	1,442	1,938	14,395	x
08-Takamura-c	GermanSentiSpin	(Takamura et al., 2005)	50,084	235	55,241	105,560	x
09-Takamura-d	GermanSentiSpin	(Takamura et al., 2005)	42,276	1	46,648	88,925	
10-Rauh		(Rauh, 2018)	17,330	0	19,750	37,080	
11-Du	SentiLitKrit	(Du and Mellmann, 2019)	1,800	0	1,820	3,620	
12-Asgari	UniSent	(Asgari et al., 2019)	656	0	728	1,384	
13-Waltinger	GermanPolarityClues	(Waltinger, 2010)	17,627	1,312	19,962	38,901	
14-Wolf	LIWC-De	(Wolf et al., 2008)	2,210	0	2,684	4,894	
15-Klinger	USAGE Sentiment Lexicon	(Klinger and Cimiano, 2014)	3,164	101	1,478	4,743	
16-Wilson	GermanSubjectivityClues	(Wilson et al., 2009)	3,336	749	5,742	9,827	
17-Mohammad	NRC Emotion Lexicon	(Mohammad and Turney, 2013)	1,550	6,728	2,339	10,617	
18-Ruppenhofer		(Ruppenhofer et al., 2017)	2,874	2,038	4,632	9,544	
19-Chen	Multilingual Sentiment Lexicon	(Chen and Skiena, 2014)	1,509	0	2,464	3,973	

Table 5: List of lexicons with information regarding the size of each polarity class of a lexicon, an abbreviation for further identification, the presence or absence of continuous sentiment values, and optionally, a lexicon name if given by the authors. X in column Cont. Val. marks that the lexicon contains continuous sentiment values.

Modifier	Baseline	f1	f1-delta	f1-combination-delta
POS with Treetagger	44.8	43.1	-1.7	-1.5
POS with Stanza	44.8	43.1	-1.8	-1.8
Stemming with Cistem	44.8	51.2	6.3	5.0
Stemming with Snowball	44.8	50.8	6.0	4.7
Lemmatization with Treetagger	44.8	50.4	5.5	5.1
Lemmatization with IWNLP	44.8	50.5	5.6	5.0
Lowercasing	44.8	47.6	2.8	0.2
Emoticons	44.8	47.8	2.9	1.7
Stop Words List	44.8	44.5	-0.3	-0.2
Valence Shifter	44.8	44.9	0.0	-0.2
Valence Intensifier and Diminusher	44.8	45.4	0.5	0.4

Table 6: Results of the modifier evaluation. Baseline is average f1 value without any modification across all corpora and lexicons. F1 the new value when only the specific modifier is added. F1-delta the difference between f1 and the baseline. F1-combination-delta is the average of all differences of all configuration with modifier turned on and off.

		Selected Modifications								
Lexicon Abbr.	f1	Lemmatization	Stemming	Cont. Sent.-Values	Emoticons	POS	Lowercasing	Stop Words Removal	Val.-Sh.	Int. & Dim.
04-Emerson	67.3	x		x	x		x	x		
01-Remus	63.6		x	x	x					x
10-Rauh	63.6		x		x	x				
13-Waltinger	63.4		x		x					
02-Clematide	61.9		x	x	x			x		
08-Takamura-c	60.6	x		x	x		x	x		
19-Chen	60.5	x			x		x	x		
18-Ruppenhofer	59.2		x		x			x		
07-Rill	56.2			x	x				x	x
16-Wilson	56.0		x		x			x		
03-Võ	54.9		x	x	x					x
15-Klinger	54.3		x		x					
14-Wolf	53.9	x			x		x		x	x
06-Siegel-m	53.8		x	x	x					
05-Siegel-p	53.4	x		x	x		x	x		
11-Du	53.2		x		x				x	x
09-Takamura-d	52.7	x			x		x	x		x
12-Asgari	52.0		x		x					x
17-Mohammad	51.6		x		x				x	x

Table 7: Best modifier combination for each lexicon sorted by f1 values. For stemming, lemmatization and POS-tagging the results with the respective best tool are shown. X in the column marks that this modifier is used.