

Towards the Automatic Sentiment Analysis of German News and Forum Documents

Andreas Lommatzsch^(✉), Florian Bütow, Danuta Ploch, and Sahin Albayrak

Technische Universität Berlin, FG AOT, Ernst-Reuter-Platz 7,
10587 Berlin, Germany
{andreas.lommatzsch,danuta.ploch,sahin.albayrak}@dai-labor.de,
florian.buetow@campus.tu-berlin.de
<http://www.aot.tu-berlin.de/>

Abstract. The fully automated sentiment analysis on large text collections is an important task in many applications scenarios. The sentiment analysis is a challenging task due to the domain-specific language style and the variety of sentiment indicators. The basis for learning powerful sentiment classifiers are annotated datasets, but for many domains and especially with non-English texts hardly any datasets exist. In order to support the development of sentiment classifiers, we have created two corpora: The first corpus is build based on German news articles. Although news articles should be objective, they often excite subjective emotions. The second corpus consists of annotated messages from a German telecommunication forum. In this paper we describe the process of creating the corpora and discuss our approach for tracing sentiment values, defining clear rules for assigning sentiments scores. Given the corpora we train classifiers that yields good classification results and establish valuable baselines for sentiment analysis. We compare the learned classification strategies and discuss how the approaches can be transferred to new scenarios.

1 Introduction

With the growing amount of textual content on the internet, the fast and efficient processing of new information is crucial in many application scenarios. Advanced machine learning approaches have been applied to address this issue. An important research topic is the automated sentiment analysis of texts. For example, press relations departments need text-mining algorithms for monitoring how institutions or companies are perceived in the media; customer relation departments need to know how the products and services offered by a company are perceived. Moreover, sentiment analysis can be used for determining engaging topics and mining the relevance of terms. Most sentiment analysis approaches use supervised machine learning algorithms or expert-defined lexicons.

The automated sentiment analysis is a challenging task due to the fact that sentiment can be expressed in several different ways and the used vocabulary highly depends on the specific scenario. Typically, the definition for positive

and negative sentiments must consider the intentions of the authors as well as the expectations of the audience (“readers”). In this paper we explain in two different scenarios how to create suitable corpora (“datasets”) enabling supervised machine learning approaches for the task of sentiment analysis based on labeled datasets. The paper describes the process of creating the corpora and the scenario-dependent criteria for annotating the texts. Subsequently, the properties of the created corpora as well as different baseline classification algorithms are discussed.

The analyzed application scenarios are as follows: First, we explain our corpus tailored for the automatic sentiment analysis of German news articles. Our dataset consists of sentences randomly extracted from German news articles and their corresponding sentiment annotations. The scenario has been defined with respect to the requirements of a press-relation department.

The second part describes the creation of a corpora consisting of messages of a German discussion forum that focuses on telecommunication related devices and services. The automated sentiment analysis has the goal to detect discontented users. The sentiment classifier enables customer relation departments to act more quickly and to improve user satisfaction.

The remaining paper is structured as follows. In Sect. 2, we discuss the challenges in creating corpora for the automated sentiment analysis and explain the scenario-specific requirements. In Sect. 3 we analyze existing corpora and their characteristics. Section 4 describes the creation of the German News corpus in detail. The section discusses the statistics and the experiences in building a classifier based on the created corpus. Section 5 describes the creation of the forum corpus in detail and explains the performance of the learned classifiers. Finally, we present a conclusion in Sect. 6 and discuss ideas for further improving the sentiment classifiers.

2 Corpus Requirements

In order to learn a sentiment classifier for texts, an appropriate dataset is required. When selecting or creating a dataset, the following aspects must be considered:

- How well do the corpus elements correspond with the content of the use case with respect to natural language and domain?
- On which layer of abstraction should the text be annotated for the use case (e.g. phrase-layer, sentence-layer, paragraph-layer, or document-layer)?
- Do the annotation values match the objectives of the scenario (e.g. positive & negative or positive & neutral & negative)?
- Which definition of a sentiment and polarity should the corpus follow?

In this paper we analyze two scenarios:

(1) We discuss the task of learning a sentiment classifier tailored to the needs of a press relationship department of major German university. The classifier should support news articles from various sources. The sentiment classification of news articles is especially challenging because news articles tend to be objective

and avoid strong emotional words. This makes the classification of news articles more difficult compared to the analysis of colloquial language used in tweets or product reviews.

(2) We discuss the task of learning a sentiment classifier tailored to analysis of the discussion boards of a big German telecommunication company. The main challenge in the scenario is that almost all messages are related to problems or questions; but users may describe problems in a friendly, objective style. Thus, the definition of the sentiment in this scenario is focused on the style and does not consider the described problem. Additional challenges consist in the colloquial language and a big number of spelling mistakes characteristic for user-generated content.

Analyzing the specific properties of both scenarios, we find that requirements with respect to sentiment analysis differ a lot.

(1) News articles are typically longer texts reporting about an event or a specific incident (“news”). The automated sentiment analysis should not only classify complete articles, but also support the exploration of subjective “hotspots” in texts. This requires a sentiment analysis of single sentences and a corpus annotated on this layer that enables us to apply supervised methods. Simply annotating whether a sentence is either neutral or subjective is insufficient because the polarity (positive or negative) of a piece of text is essential for tracking trends in the perception of topics and named entities in the media. Therefore, we propose creating a corpus with three sentiment labels: positive, negative and neutral. We favor a sentiment definition that also considers the discussed topics and that allows human annotators to imply world knowledge during the annotation process. The sentiment annotations should reflect how the topics are perceived in the society in general in order to be able to annotate topic polarity within apparently objective sentences.

(2) Messages in discussion forums tend to come in letter-like forms with salutation and closing. The main part consists of the problem description and a personal judgment. The sentiment is sometimes explicitly (e.g. “I am upset about the answer”) or indirectly expressed by rhetorical questions (“do I have to wait for a year for getting a response?”). In order to understand the sentiment, the context is important. That is why the sentiment analysis cannot be done on word level or sentence level in this scenario. The sentiment analysis should consider the complete message having a focus on the intention of the user, giving only a low weight to (often very formal) opening and closing (flowering) phrases. In order to enable the comparison of the created corpus based on forum messages with the German news corpus, we annotate the forum corpus with the same three sentiment labels: POSITIVE, NEGATIVE and NEUTRAL. Since the forum focuses on the questions and problems reported by users, we expect a small number of positive messages; but we want to use an annotation scheme similar to the German news dataset.

Both analyzed scenarios have specific challenges and requirements highly depending on the domain and the use case. Before we present the process of creating the corpora in detail, we review existing research and discuss how this work is related to the state of the art.

3 Analysis of Available Corpora

A requirement for learning sentiment classifiers are adequate corpora. In order to apply supervised machine learning approaches the corpus must provide texts annotated with a sentiment label (“class label”). Machine Learning approaches analyze the texts and extract patterns characteristic for the classes.

Corpora developed for sentiment classification highly depend on the language and the specific application scenario. There are very few freely distributed sentiment analysis corpora created from German texts.

German News Corpus. Two popular German corpora for sentiment analysis are the PRESSRELATIONS dataset [10] and the MLSA corpus [5]. The PRESSRELATIONS dataset focuses on articles regarding German political parties, making it more suitable in sentiment classification for politics. It contains 617 news articles with 1,521 annotated statements using numeric sentiment scores between -1 and 1 . The statements are usually between one and four sentences long [10].

The MLSA corpus covers a wide spectrum of topics and uses a multi-layered approach. It consists of 270 sentences, each of them annotated on the layers: sentences (layer 1), words and phrases (layer 2) and events and expressions (layer 3). Each layer has been annotated by multiple annotators, whose annotations were then combined into inter-annotator agreement measures. The sentiment definition used by MLSA is a reasonable basis for annotating objective journalistic texts, since they consider topic polarity within objective sentences.

Corpora for the Sentiment Analysis of Forum Messages. Forums allow users to discuss all types of questions. User-generated content tends to be emotional; thus forum messages are an interesting source for sentiment analysis. Ali et al. [1] analyze positive and negative opinions in a health-related forum. They annotated 607 posts using three classes (POSITIVE, NEGATIVE, and NEUTRAL). Based on the annotated corpus, lexicon-based features as well as rules-based features have been derived. The research shows that classification precision depends on the class. Domain-optimized classifiers outperform baseline algorithms such as Naive Bayes or Support Vector Machine (SVM).

Bosco et al. [3] investigate sentiments and irony in online political discussions. The research discusses how to develop corpora for mining and analyzing opinions and sentiments in social media.

Boland et al. [6] created a German text corpus for sentiment analysis that contains sentence-wise annotations of product reviews from six categories including “Mobile”, “Tablet” and “Smartphone” that all are marked as positive, negative, mixed (positive and negative) or neutral. Still, the domain differs from our scenario. Although users write their opinion about technical devices, they do not report concrete problems with them and do not ask questions. The aim of reviews is rather to assess the overall performance of the products, and in contrast to a forum message, a review is more informative and less personal.

All presented datasets do not completely fulfill the requirements of our scenarios in terms of domain and size. That is the reason for us to create new

datasets tailored for learning classifiers. The corpora are optimized for the analysis of sentiments in German news and forum texts. Nevertheless, we take the characteristics of existing datasets into account as basis for creating new datasets.

4 Creating the German News Corpus

We analyze the scenario of analyzing the sentiment of German news articles. In order to apply a machine learning approach a corpus is needed matching the specific requirements of the scenario. Since the analysis of already existing corpora has shown, that no suitable dataset for our scenario exists, we create a new corpus. We identify the following necessary steps for the annotation process which include:

- Deciding on a sentiment definition and formulating detailed annotator instructions.
- Choosing a range of possible annotation values.
- Deciding on a document source, the corpus alignment, and the classification target domains.
- Annotating the documents based on one or multiple annotator scores.

In the following paragraphs we present our sentiment definition and describe the corpus creation procedure. Finally we give a short overview of the created corpus. To avoid discrepancies between our corpus and the target classification domain, we use news articles that were previously crawled by a press review software for university-related news in German.

4.1 Sentiment Definition and Annotation Procedure

We annotate all dataset items using a 3-value scale, distinguishing between negative, neutral and positive sentences. This annotation scheme enables us to learn a classifier separating both neutral vs. polar as well as positive vs. negative sentences. This is important because the identification of polar text in journalistic articles is an especially challenging task. For the exact definition when to assign what sentiment label, we applied the categorization presented in the MLSA corpus paper [5]. Clematide et al. discern subjective and objective texts. The authors note that objectively written texts may also contain sentiments, giving the example of a sentence that talks about rising unemployment. According to the paper the sentence is negative, since a rising unemployment is normally considered as negative by readers [5].

Figure 1 shows the steps of the annotation process. We annotate randomly selected sentences from news articles. This ensures that neutral and polar sentences are annotated to result in a realistic term distribution. The occurrence of subjective and objective as well as neutral and polar sentences is the basis for learning good classifiers in the news domain. In general, polar sentences are hard to find in news articles, if only subjective statements are considered. However,

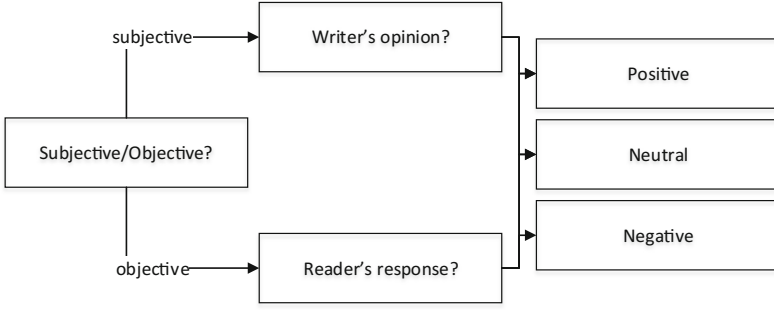


Fig. 1. The annotation procedure considers the writer’s as well as the reader’s viewpoints.

an institution or company often mentioned in sentences with negative topics (such as strikes or quality problems) is perceived negatively and might get a bad reputation. Even though, the sentences might be objective, a polarity in the perceived sentiments is interesting to find. Therefore, this should be reflected in the annotated corpus. Our sentiment definition relies on the polarity of the topic for the average reader, if the author writes in an objective style. There is another dimension to it, being the cultural or political background and personal opinions of a reader [5], which differ from person to person. For simplicity we assume there is a general consensus for many concepts and topics about their polarity.

In the first phase of the annotation, the sentences are classified by three expert annotators who all work on a separate subset of the extracted documents. After this initial phase, one of the annotators reviews the whole corpus to ensure consistency and to increase quality within the corpus. Apart from the already mentioned challenges, several discoveries were addressed in the review process.

Depending on the algorithms used for classification, a specific problem arises when annotating. Pang et al. identified sentences that mostly comprise positive words but are actually negative or vice versa [7]. Humans understand these sentences easily but they are difficult to handle by bag-of-words approaches. Such sentences may be unsuitable for optimizing the classifier for a particular scenario. In order to create a strongly tailored corpus, sentences from the text source may have to be excluded. In our case, we re-evaluated the corpus and deleted the corresponding sentences.

After the first phase of the annotation process the corpus contained only a relatively small fraction of polar sentences, because news articles are usually objective. In order to slightly attenuate the strong bias and to give the classifier additional data to judge polar sentences, additional polar sentences were added in a second annotation phase.

4.2 Corpus Statistics

The result of the annotation procedure is a corpus having a class distribution heavily leaning towards neutral sentences (see Fig. 2). With 2,369 sentences, it is also much larger than the MLSA corpus [5] and provides a source for a domain that is rarely covered by German corpora. Since the bias in the sentiment scores is created by the bias in the crawled texts, it is an interesting question whether this is actually useful for very specific classification models. From the large number of sentences it can be inferred that this bias applies to the domain of the texts as a whole.

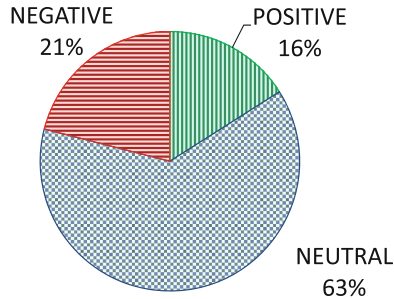


Fig. 2. German news corpus class distribution

4.3 Learning and Evaluating Sentiment Classifiers

Based on the created corpus we learn a classifier and evaluate the classifier using cross-validation.

Setup. We use the Weka machine learning framework [11] to learn a classifier based on our dataset. We choose the Multinomial Naive Bayes classifier and evaluate our model applying 10-fold cross-validation. In order to calculate the feature vectors for the sentences we use bigrams, single word tokens, a customized stop words list and the German snowball stemmer.

Results. We evaluate the classifiers using 10-fold cross-validation on the corpus. In order to get a better understanding of the performance of the Multinomial Naive Bayes classifier, we compare the strategy with a ZEROR classifier (“baseline”). ZEROR classifies every item as the class with the highest number of items in the training set. Since the neutral sentences form the majority of the sentences with 63.3%, we conduct further tests with a smaller subset of the corpus. We consider only a sample of neutral sentences to compensate for the imbalance compared to positive sentences; a class proportion of 2.5:1 is used. This does not mean the model should necessarily be created from a subset of the corpus, but it shows the effect of the bias towards the dominant class in the corpus on the evaluation results.

Table 1. Baseline comparison complete corpus [4]

	Multinomial NB			ZeroR (Baseline)		
	Negative	Neutral	Positive	Negative	Neutral	Positive
Precision	67.7%	76.3%	74.1%	-	63.3%	-
Recall	51.5%	90.3%	43.2%	0.0%	100.0%	0.0%
Accuracy	74.8%			63.3%		

Table 2. Baseline comparison 2.5:1 corpus [4]

	Multinomial NB			ZeroR (Baseline)		
	Negative	Neutral	Positive	Negative	Neutral	Positive
Precision	71.0%	70.0%	73.0%	-	52.0%	-
Recall	60.4%	85.8%	45.9%	0.0%	100.0%	0.0%
Accuracy	70.6%			52.0%		

The evaluation results (Table 1) show that baseline ZEROR classifier reaches an overall accuracy of 63.3%. This can be explained by the bias towards neutral sentences in the dataset. The Multinomial Naive Bayes classifier considerably outperforms the baseline and provides good classification results. After balancing the classes in the dataset (reducing the number of neutral sentences in the corpus), the ZEROR baseline drops (see Table 2). The classification accuracy of the Multinomial Naive Bayes drops slightly; the change is much smaller than the changes observed for the ZEROR classifier. Another interesting change is the improved recall of the polar classes with Multinomial Naive Bayes, especially of the negative class. In general, the recall of negative sentences is higher than the recall of positive sentences, with and without balancing. Possible reasons for this are the slightly smaller number of the positive instances in the corpus and the higher variance in the positive sentences.

4.4 Classifier Comparisons and Discussion

The analyzed scenario and the domain-specific bias have a big influence on the learned classifiers and the observed classification accuracy.

Optimizing the Corpus and the Learning Procedure. The use of the full corpus provides the best results in terms of accuracy. This shows that a bias has a positive influence on the overall accuracy when the same bias exists in the documents to be classified. This can be explained by the fact that the Multinomial Naive Bayes and other classifiers are more likely to classify towards that bias. If the recall of the less prominent classes is too small, one has to consider alleviating this bias. For future work, we plan analyzing whether this still applies when balancing is completely done by annotating new documents for the smaller classes instead of removing parts of the most prominent class from the corpus.

Renni et al. consider a bias based in the training set a problem in the traditional Naive Bayes classification [9]. Different applications may have different requirements in terms of recall and precision of each individual class. Thus, the bias may help improving the classification accuracy by integrating domain specific knowledge. In order to profit from the bias, the training data must be taken from the same domain or source as the test data to ensure the matching class distribution.

It could be argued that models performing well on the test set should be able to classify any sentence correctly with a good probability. Still, in our tests the overall accuracy dropped when cutting away larger parts of the instances assigned to the neutral class. This suggests that it makes sense to use a bias in the sentiment classification. Excluding sentences from the corpus would reduce the total corpus size and would require additionally annotated sentences for the smaller classes. It should also be noted that there can be other reasons to apply a bias not discussed here, one example being that certain decisions are especially costly.

Comparison with Existing Corpora. In a 3-fold cross validation with a corpus consisting of positive and negative movie reviews, Pang et al. achieved up to 81.5% classification accuracy [7]. While our results are slightly lower, we argue that there are several factors in our setup that pose new challenges [4]. Pang et al. only consider 2 classes (positive and negative reviews) in their evaluation [7]. We also consider the news domain to be at least as difficult to classify as movie reviews by the nature of the texts. News articles contain much less easily identifiable strong words that state opinions than reviews. Our sentiment definition that takes an author’s sentiment as well as the reader reaction into account may also lead to more complicated classification scenarios and therefore lower accuracy. Another difference between our setups is that their results cannot be traced back to specific hotspots in a document, since Pang et al. do not split documents to the sentence level [7]. Lastly, it has to be noted that the corpus can be used with more sophisticated classification algorithms and feature selection to improve the accuracy while still relying on the same data.

5 Creating the German Forum Corpus

In addition to the corpus based on news document, we created a corpus consisting of messages extracted from a German discussion forum focused on telecommunication topics. The forum is mainly used by persons who have problems with the telecommunication services or questions related to devices offered by a major German telecommunication company.

The sentiment analysis in the telecommunication forum is important for improving the customer satisfaction and for the efficient detection of problems with devices and services. In contrast to the news articles created by professional journalists, the messages in the forum are created by “regular” users. Thus, the analyzed messages are characterized by informal language and (partially) explicit sentiment statements.

5.1 Criteria for Labeling the Corpus

In the analyzed scenario, users typically describe concrete problems and ask other users for support or suggestions. The messages in the forum are often written in a personal, emotional style. Thus, the messages related to unpleasant problems may contain an explicit sentiment statement. Like news articles also forum messages may express sentiments implicitly. Apparently objective messages (like “Do I have to wait for a year for getting a response?”) show well the intention of the writer when including common sense for the interpretation of the message. For the creation of the Forum corpus we take both into consideration, explicit and implicit sentiments and use a three-level annotation scale:

1. We annotate messages as POSITIVE if users want to say thanks to friendly employers or praise the good service quality (e.g. increased free data volume).
2. Messages are labeled as NEUTRAL if users ask questions in an objective way (e.g. “How do I block a specific telephone number”). This means, that the fact-focused description of a problem is labeled as NEUTRAL even though the message is related to a problem.
3. We label messages as NEGATIVE, if users complain about products or services in an emotional way. Indicators for a negative sentiment are the use of emotional words, the frequent use of exclamation marks and sentences explicitly containing a negative sentiment (e.g. “I am very angry”).

The sentiment annotation considers the overall sentiment of a message. This means that salutation and complimentary close have only a very low influence on the sentiment, since even angry users may try to ensure a minimum level of politeness.

5.2 Forum and Corpus Statistics

The created corpus consists of 1,000 messages crawled from the telecommunication forum in December 2016 and January 2017. The analyzed forum is organized in threads each typically starting with a question or a problem description. The subsequent messages in the thread are comments from users and employees of the company typically providing advice or hints for tackling the described problem. Since the main use case in the analyzed scenario is the detection of arising problems, we only analyze the first message of each thread. Posts having a negative sentiment often result in “vulgar”, fruitless discussion. The automatic sentiment analysis enables the fast intervention and is the basis for improving the customer satisfaction.

In contrast to the German news corpus, we annotate the forum corpus on the message level. This is due to the observation, that typically several sentences are needed for deriving the sentiment. The messages in the dataset typically consist of 2–6 sentences (Ø5 sentences) having in average 71 words. The properties of the annotated messages are shown in Table 3.

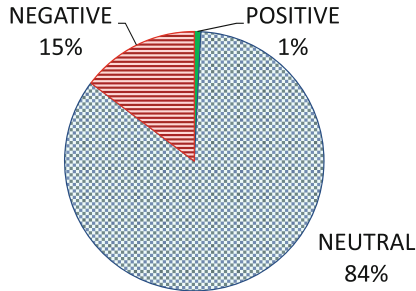
The distribution of the assigned annotations in the dataset is imbalanced. Figure 3 shows that 84% of the messages are annotated as NEUTRAL. This means

Table 3. The table lists the text message characteristics of the German Forum dataset.

Text message property	Average	Median
Number of sentences	5.4	4
Number of words	70.8	49
Number of words excluding stop words	51.9	42
Number of distinct stems (excluding stop words)	32.4	25

that the majority of analyzed forum messages are formulated in a non-emotional style. Only 1% of the messages are annotated as POSITIVE. This very small number can be explained by the fact that we only analyzed the first message of each forum thread: Only very few users start a new thread for compliment the forum employers.

Fifteen percent of the analyzed messages have a negative sentiment. The number of messages in a negative sentiment is rather small with respect to the total number of messages. This shows that the majority of users describes questions in an objective way without negative emotions.

**Fig. 3.** The class distribution in the Forum corpus

5.3 Sentiment Classification and Evaluation

We analyze different classification approaches and evaluate the classifiers based on the created corpus. At first we test classifiers based on expert-defined rules. Subsequently, we train a Multi-Nominal Naive Bayes classifier using the Weka machine learning framework.

Classification Based on Expert-Defined Rules. We analyze static, rule-based approaches tailored to the specific characteristics of the application scenario. The rules are expert-defined and not explicitly trained on the corpus.

Punctuation Mark-Based Classifier. For automatically classifying the text in the corpus, we start with a **punctuation mark-based approach**. The idea behind the approach is that emotional and (in particular) angry people

extensively use exclamation marks and sequences of questions marks. Thus, the numbers of “emotional” punctuation marks seems to be a promising indicator for identifying messages in a negative sentiment.

Due to the small number of positive messages in the created corpus, the defined classifier only considers the classes NEUTRAL and NEGATIVE. The classifier extracts all sequences of exclamation marks and questions marks. All sequences containing at least one exclamation mark are counted and weighted by the length of the sequence. In order to compute a score, the weights are summed up. If the score is above a threshold the message is classified as NEGATIVE; otherwise as NEUTRAL. We use the threshold 2 maximizing the F1 score for the class NEGATIVE.

The classification performance is presented in Table 4. The results show that sequences of exclamation marks are valuable indicators for detecting angry users; the classifier has a reasonable precision. The low recall shows that exclamation marks are only used in a small fraction of the negatively labeled messages. Overall, the exclamation mark-based classifier can be an initial starting point for the sentiment classification. In order to improve the classification accuracy it should be combined with alternative approaches.

Term-Based Classifier. In addition to the punctuation-based classification approach, we implement a classifier based expert-defined terms (“keywords”) and phrases. Therefor a set of 45 emotional words and phrases (indicating a negative sentiment) have been defined. The classifier assigns the label NEGATIVE if at least one of the words or phrases has been used in a message¹.

The classification accuracy is shown in Table 4. Overall, the classification precision is good and outperforms the punctuation mark-based classifier. Analyzing the reasons for the measured performance, we have to keep in mind, that the classifier is highly optimized for the scenario; thus the classifier tends to overfit the dataset. Moreover, the classifier relies on the occurrence of single words resulting in volatile classification results.

Nevertheless, the results show that the term-based approach tailored to the vocabulary of the specific scenario reaches a high classification accuracy.

SentiWS-Based Classifier. In order to compare the term-based approach with existing approaches, we use the SentiWS corpus [8] for building a classifier. The SentiWS corpus provides 1,650 positive entries and 1,818 negative entries. Each entry consists of a weight and the related surface forms. We classify the forum messages by computing the average sentiments score of the words (part of the SentiWS corpus). Messages having a score above a threshold (0) are classified as NEUTRAL; messages having a score below the threshold are classified

¹ Keyword/phase set: abgespeist, abzocken, ärgere, ärgerlich, arsch, beschwerde, blöd, desaster, die nase voll, dumm, dümm, enttäuscht, ex-kunde, frechheit, frustriert, grottig, hinhalten, hohn, idiot, kann doch nicht so schwer sein, katastrophe, minderwertigste, nervt, nicht kapiert, opfer, rausnehmen, reklamiert, schämen, schnauze, scheiss, schlimmer, schuld, teufel, unfassbar, unzufrieden, verärgert, vergebens, versagt, verschlimmbesserung, verschonen, verschont, vertrösten, verzweifeln, wird mir schlecht.

as negative. The threshold 0 maximizes the F1 score for the class NEGATIVE. Due to the very small number of messages labeled as POSITIVE, we implement the classifier so that only the labels neutral and negative are predicted.

The evaluation results show a lower performance of the classifier (with respect to both accuracy and recall of negative messages) compared with the other classifiers (cf. Table 4). Reasons for the low performance are that the SentiWS corpus does not match the specific properties of the forum. Moreover, the users in the forum often use polite salutations even if the message has a negative sentiment. Averaging the sentiment scores over the complete message does not seem to be an adequate approach in the forum scenario.

Table 4. Comparison of the different static classifiers

	Keyword-based		Punctuation mark-based		SentiWS-based	
	Negative	Neutral	Negative	Neutral	Negative	Neutral
Precision	62.7%	96.1%	50.0%	87.6%	21.6%	87.5%
Recall	62.7%	97.0%	27.2%	94.5%	50.3%	67.4%
Accuracy	93.7%		84.6%		64.9%	

Classification Applying Machine Learning. We use the Weka machine learning framework [11] to learn a classifier based on our dataset. We choose the Multi-nominal Naive Bayes classifier and evaluate our model applying 10-fold cross-validation. In order to calculate the feature vectors we use single word tokens, a default German stop word set² and the German snowball stemmer.

Results. The results (cf. Table 5) have been achieved using 10-fold cross-validation on the corpus. The Naive Bayes Classifier performs only slightly better than the baseline recommender (ZEROR). This shows that single words are only weak features for predicting the sentiment. We analyze the terms that have the highest impact in the learned classifier. The terms having the highest weight for identifying messages labeled as NEGATIVE are: “hotline”, “wait”, “week”, and “long”. This indicates that messages with a negative sentiment are often related to long response times (users must wait for weeks) and to problems with the hotline.

Nevertheless, the sentiment detection is challenging. Most messages are written in an objective style. Even messages labeled as NEGATIVE often contain a neutral problem description. Thus, the analysis on message level might not give enough weight to the emotional sentences. An annotation on sentence level could address this issue.

Another challenge in the analyzed dataset is the strong bias towards messages labeled as NEUTRAL. Due to the limited size of the dataset, the number of training samples for the rare classes (POSITIVE, NEGATIVE) is rather small. In order to address this issue, we will extend the corpus with additional messages.

² c.f. https://lucene.apache.org/core/6_2_0/analyzers-common.

Table 5. Comparison of the classification strategies learned on the Forum corpus evaluated using 10-fold cross-validation.

	Multinomial Naive Bayes			ZeroR (Baseline)		
	Negative	Neutral	Positive	Negative	Neutral	Positive
Precision	60.3%	91.0%	33.3%	-	63.3%	-
Recall	49.7%	94.3%	12.5%	0.0%	100.0%	0.0%
Accuracy	87.1%			84.5%		

5.4 Discussion

The comparison of the different classification approaches shows that the sentiment classification is a challenging task. Overall, the term-based classifiers performed best. Due to the specific characteristics of the scenario, the term-based classifiers must be optimized to the analyzed use-case. Classifiers build based on documents from other domains (e.g. the SentiWS classifier) show a significantly lower accuracy.

The presented German sentiment forum dataset has a strong bias towards NEUTRAL messages. The number of messages having a positive sentiment is so small that not sufficient training examples exist for learning rules how to identify positive messages.

Overall, the most messages in the analyzed forum are written in a polite, objective style. The automatic detection of messages having a NEGATIVE sentiment is hard. Since the dataset is labeled on message layer, the extraction of the discriminating terms and phrases is an additional challenge. Due to the specific vocabulary and user habits in the forum, classifiers must be optimized. In the analyzed use case the detection of upset, annoyed users is one important task (measured based on the recall for the class NEGATIVE). The analyzed classifiers show, that keyword-based algorithms are a promising starting point for further improvements.

As future work, increasing the corpus size, additional text analysis methods as well as more complex classification algorithms (based on artificial neural networks or deep learning approaches) should be researched. In addition, the combination of different classifiers is a promising an approach. Since negative sentiments can be expressed in several different ways, an ensemble of classifiers, each analyzing one specific aspect, should be able to reach a significantly higher accuracy than a single classifier.

6 Conclusion and Future Work

In this paper we presented the creation of two new corpora tailored to the sentiment analysis of German news articles as well as German forum messages. We discussed the relevant challenges and solutions when creating a dataset for learning a sentiment classifier. In addition, we explained the characteristics of our

datasets and showed how to learn powerful classifiers based on the datasets. Since sentiment analysis corpora for languages other than English are very rare, this provides new opportunities to increase the classification quality in the domains at hand. We have discussed the challenges that occur and have considered solutions and open questions to be answered in the future. We use the created corpus to classify documents with good results. Many papers focus on tweaking classification algorithms and feature selection, but we consider the corpus creation to be an important step as well. Sentiment definition and bias can have a significant impact on the classification. The created datasets show that there is often a strong bias towards one class. A strong bias makes the interpretation of accuracy values difficult due to the fact that classifiers always predicting the most dominant (“most popular”) class reach a high accuracy.

One interesting topic for the future research is the cost-sensitive learning. The consideration of costs could counteract the bias inherent in the datasets. We argue that there is no strict necessity to remove a bias since it can be useful if the target domain exhibits it as well, or if the use case demands it. Another related research topic would be the consideration of a specific machine learning approach when creating a corpus. Considering our examination of the MLSA corpus, enhancements in the form of inter-annotator agreements could also be applied in the future to further improve the annotation quality. Finally, the most important results of our work are the corpora itself as one of the few German sentiment analysis datasets. They can be used as a starting point to improve classification by focusing on the algorithms instead, or as a basis for bigger corpora.

Another interesting research question is how sentiment classifiers learned for a specific domain can be transferred to other domains. Our analysis shows that classifiers optimized for one application scenario often perform poorly when using these models in new use cases. Balahur et al. state that the sentiment of news articles is a unique domain with special needs [2]. They propose that for news sentiment analysis, the source, the target and different perspectives on an article (reader interpretation, author intention) should be considered. As we use Multi-nominal Naive Bayes, identifying source, target and different perspectives would be possible for the annotators but would not make a difference for the algorithm. When working with more sophisticated algorithms and feature selection methods, going back to these observations may help in further improving sentiment analysis.

Acknowledgment. This work was supported in part by the German Federal Ministry of Education and Research (BMBF) under the grant number 01IS16046.

References

1. Ali, T., Schramm, D., Sokolova, M., Inkpen, D.: Can I hear you? Sentiment analysis on medical forums. In: Proceedings of the International Joint Conference on Natural Language Processing 2013, pp. 667–673. ACL (2013)

2. Balahur, A., Steinberger, R.: Rethinking sentiment analysis in the news: from theory to practice and back. In: *Proceeding of WOMSA*, vol. 9 (2009)
3. Bosco, C., Patti, V., Bolioli, A.: Developing corpora for sentiment analysis: the case of irony and Senti-TUT. *IEEE Intell. Syst.* **28**(2), 55–63 (2013)
4. Bütow, F., Schultze, F., Strauch, L., Ploch, D., Lommatzsch, A.: Sentiment analysis with machine learning algorithms on German news articles. Project report, Berlin Institute of Technology, AOT (2015). <http://www.dai-labor.de/publikationen/1052>
5. Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U., Wiegand, M.: MLSA-A multi-layered reference corpus for German sentiment analysis. In: *LREC*, pp. 3551–3556 (2012)
6. Boland, K., Wira-Alam, A., Messerschmidt, R.: Creating an annotated corpus for sentiment analysis of German product reviews. Monograph, GESIS - Leibniz-Institut für Sozialwissenschaften (2013). http://www.ssoar.info/ssoar/bitstream/handle/document/33939/ssoar-2013-boland.et.al-Creating.an.Annotated.Corpus_for.pdf
7. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
8. Remus, R., Quasthoff, U., Heyer, G.: SentiWS - a publicly available German-language resource for sentiment analysis. In: Chair, N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA), Valletta, Malta (2010)
9. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive Bayes text classifiers. In: *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, vol. 3, pp. 616–623 (2003)
10. Scholz, T., Conrad, S., Hillekamps, L.: Opinion mining on a German corpus of a media response analysis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) *TSD 2012. LNCS*, vol. 7499, pp. 39–46. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-32790-2_4](https://doi.org/10.1007/978-3-642-32790-2_4)
11. University of Waikato: Weka 3 - Data Mining with Open Source Machine Learning Software in Java. <http://www.cs.waikato.ac.nz/ml/weka>

Innovations for Community Services

17th International Conference, I4CS 2017, Darmstadt,

Germany, June 26-28, 2017, Proceedings

Eichler, G.; Erfurth, C.; Fahrnberger, G. (Eds.)

2017, XII, 197 p. 63 illus., Softcover

ISBN: 978-3-319-60446-6