Creation of a German Corpus for Internet News Sentiment Analysis

Florian Bütow, Andreas Lommatzsch, and Danuta Ploch

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

Abstract. The fully automated sentiment analysis on large text collections is an important task in many applications. It is often solved applying supervised machine learning algorithms. The basis for learning powerful sentiment classifiers are annotated datasets, but for many domains and non-English texts hardly or even no datasets exist. In order to support the development of sentiment classifiers for German news articles, we create a new corpus of annotated German news articles related to the Berlin Institute of Technology. Although news articles should be objective, they often excite subjective emotions. In this paper we describe the process of creating a corpus for news documents and discuss our approach for tracing sentiment values back to its hotspots, defining clear rules for assigning polarity scores, and handling the imbalance of class labels. The created corpus consists of sentences each labeled either as NEUTRAL, POSITIVE or NEGATIVE. Given the corpus we train a classifier that yields good classification results and establishes a valuable baseline for sentiment analysis on German news articles.

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. E.g., press relations departments need text-mining algorithms for monitoring how institutions or companies are perceived in the media. 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.

In this paper, we focus on a supervised machine learning approach: We create a new, manually annotated dataset and train a classifier. Our dataset consists of sentences randomly extracted from German news articles and their corresponding sentiment annotations. We explain the process of creating the corpus and discuss the characteristics of the corpus. The remaining paper is structured

as follows. In Sec. 2, we discuss the challenges in creating corpora for the automatized sentiment analysis and explain the scenario-specific requirements. In Sec. 3 we analyze existing corpora and their characteristics. Sec. 4 describes the creation of our corpus in detail. Dataset statistics and the experiences in learning a classifier based on the created corpus follow in Sec. 5. Finally, we present a conclusion in Sec. 6 and discuss ideas for further improving the sentiment classifier.

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 does the content of the corpus correspond with the content of the use case in terms of natural language and domain?
- On which layer the text should be annotated for the use case (e.g. phrase-layer, sentence-layer, paragraph-layer)?
- Is the range of annotated values sufficient for the goal (e.g. positive & negative or positive & neutral & negative)?
- Which definition of a sentiment and polarity should the corpus follow?

We analyze the task of learning a sentiment classifier tailored to the needs of classifying German news articles related to a major German university. The sentiment classification of news articles is especially challenging because news articles tend to be objective and to avoid strong emotional words. This makes the classification of news articles more difficult compared to the analysis of colloquial language used in tweets or reviews.In addition, most existing sentiment approaches and available corpora focus on English texts.

Sentiment analysis approaches 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 that layer that enables us applying 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.

3 Analysis of Available Corpora

There are very few freely distributed sentiment analysis corpora created from German texts. Two popular corpora are the PRESSRELATIONS dataset [6] and the

MLSA corpus [3]. The PRESS-RELATIONS 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 [6].

The MLSA corpus covers a wide spectrum of topics and uses a multi-layered approach. Each layer has been annotated by multiple annotators, whose annotations were then combined into two different 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.

Both presented datasets do not fulfill completely our requirements. Nevertheless, we use the characteristics of these datasets as basis for creating a new dataset. For our dataset we use sentence-layer annotations. Our annotation scheme is strongly influenced by the sentiment annotation scheme used in the MLSA corpus. In contrast to the MLSA corpus we use news articles related to universities. Moreover, we decided to create a larger corpus (compared with MLSA) for improving the classification accuracy and the significance of the evaluation results.

4 Creating the Corpus

Since the already existing corpora are not suitable for our scenario 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 the problem of 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 & 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

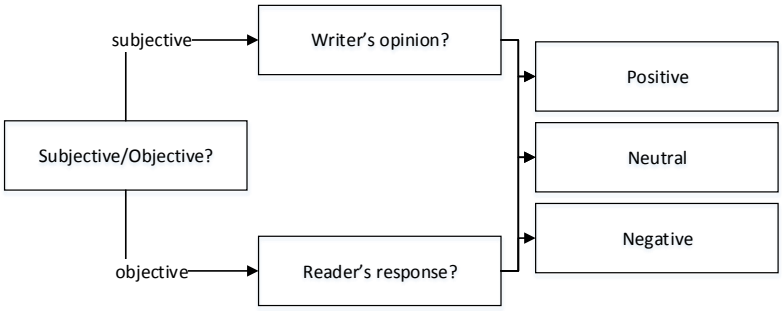

Fig. 1. The annotation procedure considers the writer's as well as the reader's viewpoints.

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 [3]. 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 [3]. Fig. 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 for ensuring 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, 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. The process for the annotation of a sentence shown in Fig. 1. 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 [3], which might 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 [4]. 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 are 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 [3] and provides a source for a domain that is hardly covered by such 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.

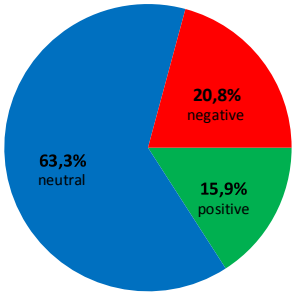

Fig. 2. Corpus class distribution

5 Classification

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

5.1 Setup

We use the Weka machine learning framework [7] 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 for the sentences we use bigrams, single word tokens, a customized stop words list and the German snowball stemmer.

5.2 Results

Table 1. Baseline comparison complete corpus [2]

	Multinominal NB			ZeroR (Baseline)		
	Negative	Neutral	Positive	Negative	Neutral	Positive
Precision	67.7 %	76.3 %	74.1 %	0.0 %	63.3 %	0.0 %
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 [2]

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

The results have been achieved using 10-fold cross-validation on the corpus. In order to get a better understanding of the performance of the Multi-nominal 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 apply a balancing of the most dominant class (neutral) to the least prominent class (positive) in a 2.5:1 relation. This does not mean the model should necessarily be created from a subset of the corpus, but it shows the advantages of the corpus compared to the baseline.

Using the full corpus, Multi-nominal Naive Bayes provides good results, with ZeroR being only slightly behind it in terms of overall accuracy (see Table 1). After balancing, with fewer neutral sentences in the corpus, the ZeroR baseline drops (see Table 2). The classification accuracy of the Multi-nominal 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 Multi-nominal 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.

5.3 Discussion

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 Multi-nominal 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 [5]. 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 for ensuring 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.

5.4 Classification Comparisons

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 [4]. While our results are slightly lower, we argue that there are several factors in our setup that pose new challenges [2]. Pang et al. only consider 2 classes (positive and negative reviews) in their evaluation [4]. 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. don't split documents to the sentence level [4]. 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.

6 Conclusion and Future work

In this paper we presented the creation of a new corpus tailored to the sentiment analysis of German news articles. We discussed the relevant challenges and solutions when creating a dataset for learning a sentiment classifier. In addition, we explained the characteristics of our dataset and showed how to learn powerful classifiers based on the dataset. Since sentiment analysis corpora for languages other than English are very rare, this provides new opportunities to increase the classification quality in this domain. 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 of almost 75% accuracy in our press review scenario. 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. We showed advantages and disadvantages of balancing the corpus. Because the sentiment definition for a corpus and the classification goals must match, it is often a challenge to find an appropriate corpus that is annotated with the same definition in mind.

One interesting topic for the future is the comparison of the results from a biased corpus with the results from a bias by weighted costs for the classes. We argue that there is no strict necessity to remove a bias since it can be useful if the target domain corresponds with it, or if the use case demands it. Another related research topic would be the consideration of a specific classifier when creating a corpus and how this could influence the results. 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 result of our work is the corpus itself as one of the few German sentiment analysis datasets. It can be used as a starting point to improve classification by focusing on the algorithms instead, or even as basis for a bigger corpus.

Another interesting research question is discussed by Balahur et al. who consider news articles as an unique domain for sentiment analysis with special needs [1]. 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 however, going back to these observations may help in further improving news sentiment analysis.

References

1. Balahur, A., Steinberger, R.: Rethinking sentiment analysis in the news: from theory to practice and back. Proceeding of WOMSA 9 (2009)
2. 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
3. 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)
4. 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)
5. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R., et al.: 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)
6. Scholz, T., Conrad, S., Hillekamps, L.: Opinion mining on a german corpus of a media response analysis. In: Text, Speech and Dialogue. pp. 39–46. Springer (2012)
7. University of Waikato: Weka 3 - Data Mining with Open Source Machine Learning Software in Java. website, available at `http://www.cs.waikato.ac.nz/ml/weka`