Sentiment-Analysis for German Employer Reviews

Jennifer Abel, Katharina Klohs, Holger Lehmann,
and Birger Lantow(✉)

Department of Business Information Systems,
Faculty of Computer Science and Electrical Engineering,
The University of Rostock, Albert-Einstein-Str. 22, 18051 Rostock, Germany
{jennifer.abel, katharina.klohs, holger.lehmann,
birger.lantow}@uni-rostock.de

Abstract. This paper examines the possibilities of sentiment analysis performed on German employer reviews. In times of competition for highly skilled professionals on the German job market, there is a demand for the monitoring of social media and web sites providing employment related information. Compared to mainstream research this implies (1) a focus on German language, (2) employer reputation as a new domain, and (3) employer reviews as a new source possibly showing special linguistic characteristics. General approaches and tools for sentiment analysis and their application to German language are assessed in a first step. Then, selected approaches are evaluated regarding their analysis accuracy based on a data set containing German employer reviews. The results are used to conclude major obstacles, promising approaches and possible prospective research directions in the domain of employer reputation analysis.

Keywords: Sentiment analysis · Recruitment · Social media analysis · Employer reputation · Machine learning

1 Introduction

In 2016, 70% of German information technology companies stated that there is a lack regarding professionals on the job market[1]. The situation in other German industry branches is similar. Thus, there is a competition for skilled workforce on the job market. In consequence, the concepts of social media marketing are approaching human resources departments. The volume of user-generated content in social networks (Twitter or Facebook) as well as in Blogs increases continuously[2]. Additionally, the market of specialized portals that provide employer ratings based on user generated content is growing as well.

[1] https://www.bitkom.org/Presse/Pressegrafik/2016/November/Bitkom-Charts-IT-Fachkraefte-14-11-2016-final.pdf.

[2] http://epceurope.eu/wp-content/uploads/2015/09/epc-trends-social-media.pdf.

W. Abramowicz (Ed.): BIS 2017 Workshops, LNBIP 303, pp. 37–48, 2017.
https://doi.org/10.1007/978-3-319-69023-0_4

Web portals like Kununu[3], Jobvoting[4], or Glassdoor[5] offer a platform for writing reviews about employers. In the course of human resources management, the monitoring of the attitude (the sentiment) towards the employer in these portals and also in the rest of the social web is an important element, which is usually controlled manually nowadays.

Next to the evaluation of (semi-)structured data from employer-review-specialized portals, there is also a need of an analysis of unstructured text-data from other sources like Facebook and Twitter. Considering the rising data volume[6], an automation of this process is desired. Here, sentiment analysis plays a major role. Sentiment analysis classifies human communication content regarding positive or negative sentiment [1]. This can be generally applied to all kinds of human communication. In the context of our work we consider textual data only. The research in this area has progressed in recent years [2]. Especially for text mining, some procedures are established and a lot of tools are available, which can be used for the task of sentiment analysis.

The goal of this paper is to examine the possibilities for an analysis of German employer reviews with reference to existing sentiment-analysis approaches. This represents a transfer to a new domain compared to known approaches like assessments of movies or products, and evaluations of brands. Besides the domain related content, a different language style in this domain may lead to different results compared to earlier research. Furthermore, there is a transfer regarding the used language. Compared to German language, English is spoken by a lot more people and shows a simpler grammar. Worldwide, English is spoken by 1.5 billion people compared to only 185 million people that speak German[7]. Thus, research and tool support mainly focuses on English language. The following research questions addressed in our paper:

1. What kind of effort is necessary to apply sentiment analysis methods to a certain language (German) and domain (Employer Reputation)?
2. How well do the selected approaches evaluate German employer ratings in comparison?
3. What conclusions can be generally drawn for the selected domain?

Generally, this study has an explorative character. Thus, the observed phenomena and problems are in focus and not the significance of differences in performance or the estimation of effort. In order to answer the research questions, this work is structured as follows. In the next section, basic approaches for sentiment analysis are briefly explained. This serves for the classification and description of available tools, and for the identification of required language specific artefacts. This is followed by an experiment (Sect. 3) and a summary together with the outlook in Sect. 4.

[3] www.kununu.com.

[4] www.jobvoting.de.

[5] www.glassdoor.com.

[6] http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

[7] https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/.

2 Sentiment Analysis Approaches

In order to classify approaches for sentiment analysis, a distinction is made regarding the object of analysis (classification level) and the used technology.

According to [3, 11] sentiment analysis approaches can be divided in the following classification levels: (1) word, (2) sentence, and (3) document. The word-level inspects individual words for their polarity and usually distinguishes between positive, negative and neutral words. At the sentence-level the polarity of each individual sentence is assessed by aggregating the words' polarities. At document level sentence polarities are aggregated for a complete document. Several rules can be applied for that aggregation. The assessment on word or sentence level can identify different sentiments in different parts of a larger structure (sentence or document).

Regarding the used technology, two main approaches for sentiment analysis exist [1]: Machine learning and the lexicon-based approach. A third technique is the hybrid approach, which is a combination of machine learning and lexicon-based techniques. This is mentioned here for completeness, but not further explained.

2.1 Machine Learning

Machine learning approaches can generally be divided into supervised and unsupervised learning. In the case of supervised learning, the goal is to learn a mapping of input values to output values while the correct output values are known [4]. This is done using a training data set. A learning algorithm learns a model by minimizing the mapping bias compared to the correct mapping provided by the training data set [5]. For the selected application domain, this approach means the generation of appropriate training data by collecting German-language employer evaluations and classifying them manually with regard to their sentiment.

Prior to the application of machine learning, generally a pre-processing of the texts is required in order to perform feature selection and feature extraction [11]. A feature in this context is an attribute (e.g. frequency of a certain term) whose value is used for classification. Pre-processing typically includes tokenization, stemming/lemmatizing, filtering stop-words, and pruning for feature selection, resulting in a word vector. For feature extraction, a measure for the importance of the single words in the vector for each text/document has to be calculated. The *Term Frequency-Inverse Document Frequency* (TF-IDF) is a metric that multiplies the two quantities Term Frequency (TF, number of term occurances) and Inverse Document Frequency (IDF, inverse of the number of documents that contain the term) and has proven to be well fitted for text classification [7]. The TF-IDF-measure is used for the evaluation in Sect. 3. For these steps, there is a good tool support. However, setting the parameters for text-preprocessing can have a large influence on the performance of machine learning technologies (see Sect. 3).

In the area of supervised learning, there are a variety of usable technologies. Popular and often supported approaches are *Bayes Classifiers*, *Support Vector Machines* (*SVM*), *Decision Trees*, and *Artificial Neural Networks* (see also [8]). The *Bayes Classification* is based on the Bayes' law, which can be used to calculate the conditional probability of an event under a particular condition. The classifier represents each object by an n-dimensional vector, where n is the number of features of an object. Accordingly, an object is assigned to the class it belongs to with the highest probability [9]. *SVMs* construct a plane which separates the instances of the classes in the feature space as best as possible. Such planes, however, are usually in a multi-dimensional space. Thus, their determination can cause large computational effort [8]. *Decision Trees* are a widespread method, which is used in many areas. The CART method, the CHAID method or the ID3 algorithm are the most common forms [10]. The advantage of *Decision Trees* is the simplicity in understanding and interpreting the results [13]. However, they are not well fitted for classification tasks with a high number of features such as natural language texts. *Artificial Neural Networks* are information-processing systems whose structure and mode of operation are similar to those of nervous systems, especially to the brain of animals and humans. There is the possibility of supervised as well as unsupervised learning. *Artificial Neural Networks* are characterized by a high computational effort.

In the case of unsupervised learning, the output values are not known and only the input values are available. Therefore, it is necessary to determine regularities in the input values [6]. These learning methods are generally unsuitable for the chosen application, because the resulting clusters do not need to have any relation to the expressed sentiment.

2.2 Lexicon-Based Analysis

Ravi and Ravi describe in [11] a lexicon as a sentiment vocabulary. Thus, the elements of a vocabulary are tagged with a polarity (positive or negative), as well as a degree of strength. The generation of a lexicon starts with a seed word, which is tested for synonyms and antonyms by means of a lexicon such as WordNet. Based on the seed, polarity and its strength can be derived for synonyms and antonyms. The idea of the lexicon-based sentiment analysis is to search for positive and negative words within a sentence and to use these occurrences for sentence classification. This should also include the sentiment strength. Sentences like: "*This book is good.*" and "*This book is excellent.*" could be not distinguished regarding sentiment otherwise. Therefore, strength values are used for weighting the single word sentiments. This rule cannot be applied straight forward because there are constructs like negations or amplification words. Reinforcing words such as "*very, a little, quite, ...*" have no positive or negative polarity, but they reinforce or weaken the following words in a sentence in their sentiment. The use of a negation in a sentence can lead to a negation of the complete sentence. Individual sentences containing both positive and negative parts are also problematic. These are predominantly sentences with binding words such as "*but, although*" [12]. These issues are addressed by rules that base on sentence and terms structure.

2.3 Problems of Sentiment-Analysis

There may occur various problems using the mentioned methods of sentiment analysis. A major problem of the lexicon-based approach is that it cannot be predicted how the sentiment is expressed. For example, it is difficult to define the sentiment orientation of domain specific expressions. Thus, a “long battery lifetime” is considered to be positive, but a “long waiting time” is considered negative [15]. However, a context or domain specific lexicon may help to solve this problem if the context is known.

In the case of machine learning approaches, the problem is the model fitting to a certain domain. If a system has been trained with a corpus of movie reviews, it will deliver less accurate results with regard to, for example, employer evaluations. A solution would be to train the model for all possible application domains [16]. However, this increases effort. Both methods have limited performance regarding the detection of irony and sarcasm. Even humans have difficulties here. A lot of context information has to be considered and interpreted correctly in order to detect these phenomena. For example, the simple statement “This is awesome.” can be connoted positive but also negative as a sarcastic statement [17].

3 Experimental Evaluation of Analysis Methods

This study focuses on the effort for adapting or creating classifiers for the given domain and language and on the accuracy of sentiment analysis. This addresses research questions 1 and 3 presented in the introduction.

Although the used tools may have an influence on the results, the general process of implementing one of the analysis approaches for a certain application domain is tool independent as well as general assumptions regarding classification accuracy. Section 3.1 briefly describes the tool selection process and its results. The data set used for the assessment is discussed in Sect. 3.2. The actual execution of the experiment is then described (Sect. 3.3). At last, the results are evaluated (Sect. 3.4).

3.1 Sentiment Analysis Tool Selection

The search for appropriate sentiment analysis tools was based on the surveys by Ravi and Ravi [11] and Pang and Lee [18]. Additionally, a Google-search for German sentiment analysis tools has been conducted. Overall 35 tools have been found. 17 were able to process German texts. Only two met the criteria (a) availability as open source and (b) German language support. Thus, RapidMiner has been selected for the evaluation of machine learning based approaches and SentiStrength for lexicon-based approaches.

RapidMiner is an open-source data-mining-tool. It offers many different analytical methods, including simple statistical evaluations, correlation and regression analyses, classification and clustering [19]. Besides operators for data extraction and text analysis (tokenizer, stemmer etc.) there are operators provided that implemented machine learning algorithms such as *Decision Tree*, *Artificial Neural Nets*, *SVM*, and *Naïve Bayes.* An analysis at sentence and document level is possible [19].

SentiStrength is a lexicon-based application for sentiment-analysis of short texts. It was originally developed for English language. The aim of sentiment analysis with SentiStrength is to calculate a cumulative document polarity based on the polarity of single words and expressions [25]. Thus, in the lexicon, there is a word list associated with positive and negative polarity values. The result of an analysis is the strength of negative and positive polarity in a given document [21, 25]. According to [25] two scales are used, because psychological research showed that people process positive and negative sentiments at the same time. This can occur especially when a person has mixed feelings about a situation and weighs the advantages and disadvantages. Using SentiStrength, a text has a positive polarity if the analysis value is between 2 and 5 (negative between −2 and −5) [21]. Accordingly, the values 1 and −1 are neutral and are ignored in the calculation.

The lexicon originally came from manually annotated MySpace comments regarding positive/negative polarity [25]. Later, the lexicon resource has been extended with annotated data of the General Inquirer Lexicon[8] (LIWC) [15]. SentiStrength supports rules for negation and amplification of sentiments. The German version of SentiStrength was developed by Pirker and Kyewsk [21]. It is based on a translation of the LIWC data in 2011 using the Google translator. Furthermore, manual corrections and additions have been made. A similar process for creating a language specific lexicon is described by Momtazi in [12] for German language and by Cirqueira et al. for Portuguese language in [24].

3.2 Example Data Set

In order to check the correctness of sentiment analysis, an annotated data set for training and/or comparison is needed (see Sect. 2). Manual annotation is very time-consuming and contradicts the concept of automation behind the analysis approaches [22].

Thus, the employer evaluation portal Kununu has been selected as the data source. Besides a textual comment, users can rate the employers by a five-star rating scale. These ratings on a Likert scale can be used for automated annotation/classification regarding the sentiment and reduces the effort for manual coding. In contrast to the majority of social media content, comments on Kununu are rather long and contain less colloquial language. Emoticons and letter repetitions are seldom used, which therefore do not have to be considered further in automated processing. This reflects the domain specific language style.

A training data set of 1200 records was collected manually. Thelwall et al. recommend in [21] a size of 2000 records for training and evaluation. As shown later in Sect. 3.3, this number does not need to be matched to draw conclusions. Although a larger data set results in better outcomes, this effect is small. The following criteria have been applied for record selection: (1) Even distribution of comment types (pos./neg.) per employer, (2) German language comments, (3) Clear polarity regarding the star rating (see below). On average the use of 28 comments per company was noticed.

In order to classify the data regarding positive/negative polarity, the available star ratings on the basis of a Likert scale were used. The individual comments were classified as positive if the given number of stars was greater than or equal to 4. A negative classification was carried out at less than 2.5 stars. The discrepancy in the selection results from the fact that it is not possible to assign 0 stars (the minimum is 1). In

[8] http://liwc.wpengine.com/.

addition, significantly more comments were positively evaluated by users. The comments having a rating between 2.5 and 4 stars were not included in the training data set to draw a clear boundary between positive and negative.

As recommended by [12], an additional test data set of 120 records (equally distributed positive/negative, 10% of the training data size) has been used for the evaluation of classification accuracy.

Using machine learning, the employer name within the comments may be determined by the learning algorithm as the best discriminator for positive or negative polarity. Hence, a certain company is a "bad" employer and thus all comments including its name are negative. To avoid this overfitting of the learned model, company names have been used as stop-words and are thus filtered out for sentiment analysis. This filtering is generally not required for lexicon based approaches.

3.3 Implementation of the Experiment

As already described, several learning algorithms are available in RapidMiner. For the evaluation and comparison of the automated classification, *Naïve Bayes* and *SVM* have been selected. *Naïve Bayes* is a very efficient approach which can perform very well in certain situations while the *SVM* is outstanding regarding accuracy [13]. Considering other approaches supported by RapidMiner - decision trees show a poor performance for text analysis and artificial neural network consume a lot computing resources [13].

The pre-processing of texts was done using standard RapidMiner components. Figure 1 shows the used process. For stemming, which needs a language specific algorithm, RapidMiner implements the stemmer by Caumanns presented in [14]. The classification accuracy had a large sensitivity regarding the lower bound for the pruning step (removing infrequent terms). Including words that occurred in less than 0.5% of the comments resulted in a drop in classification accuracy. A lot of these words was equally distributed in positive and negative comments, but was used by the machine learners to determine positive polarity. The reason was the higher term frequency (TF) in positive comments due to their generally lower length. An optimum for the lower bound was manually found at 2%. Thus, all words that occurred in less than 2% of the comments have been excluded from the word vector.

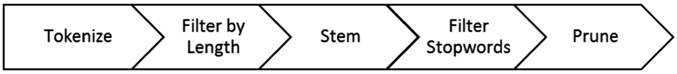

Fig. 1. Text pre-processing in RapidMiner

The provision of training data is a main driver for the effort caused by supervised machine learning. Thus, the influence of different training data set sizes on classification accuracy was investigated by learning the *SVM* and the *Naïve Bayes* classifiers with 300, 600, 900 and last 1200 records of the available training data. Furthermore, the size of the used word vector which equals to the number of features used for classification was monitored. If there is a saturation regarding the number of features, an increase of the training data set size will have little effect on accuracy.

In order to map the 2 scale rating of SentiStrength to a positive/negative classification, all comments that showed a negative polarity have been classified negative, regardless of the value in the positive scale. SentiStrength as a lexicon based tool does not need a training. Thus, the effect of the training data set size has not been evaluated. However, for the context of the given domain there are specific words that have a positive or negative polarity which is not the case in their general usage. The problem of context-relatedness has already been discussed in Sect. 2.3. For example, “pressure” is negative in a work context and “training” is positive. Thus, the effect of collecting such terms from a set of 60 comments training data and adding them to the lexicon in order to create a context specific lexicon has been evaluated. There was no change in the accuracy. Although there was a certain qualitative saturation (newly assessed comments did not provide data for new lexicon entries), the new words in the lexicon have not been found in the test data.

The accuracy of the classification results is examined by means of the three evaluation measures Recall, Precision and F-Measure [26]. They are commonly used in data analysis [23, 27–30]. The values are calculated regarding the negative polarity classification because the determination of negative comments is important in the given application domain. Calculation is based on numbers of correctly (true positive - tp and true negative - tn) and falsely (false negative - fn and false positive - fp) categorized comments. The formulas for determining the negative polarity class recall and precision values are listed below.

$$Recall = \frac{tn}{fp + tn} \quad Precision = \frac{tn}{tn + fn} \quad F\!-\!Measure = \frac{2 * R * P}{P + R} \tag{1}$$

The recall measures how many of the actually negative comments in the database were found in relation to the number of all negative comments in the database. The precision is the ratio of the correctly negatively classified comments to the total number of all negatively classified comments. The F-Measure is the harmonic mean of recall and the precision. It serves as a general measure for classification performance. Figure 2 shows the resulting values for the different training set sizes. Table 1 compares the tested methods with their obtained values for Precision, Recall and F-Measure regarding negative polarity in percent.

3.4 Evaluation of Experimental Results

Looking at research question 1 - What kind of effort is necessary for an analysis of German-language texts? – a first answer is that a lexicon based approach can reach a good performance without any effort regarding the creation of a training data set or a lexicon, as long as a general language specific sentiment lexicon is available. When it comes to machine learning, a next question comes up regarding the effort - How does the size of the training data set influence the accuracy of the machine learning approaches?

There is a stable performance of *SVM* around 80% F-Measure. The *Naïve Bayes* accuracy in contrast increases with an increasing training data size but does not match the SVM performance. Based on these conclusions and the other experimental results,

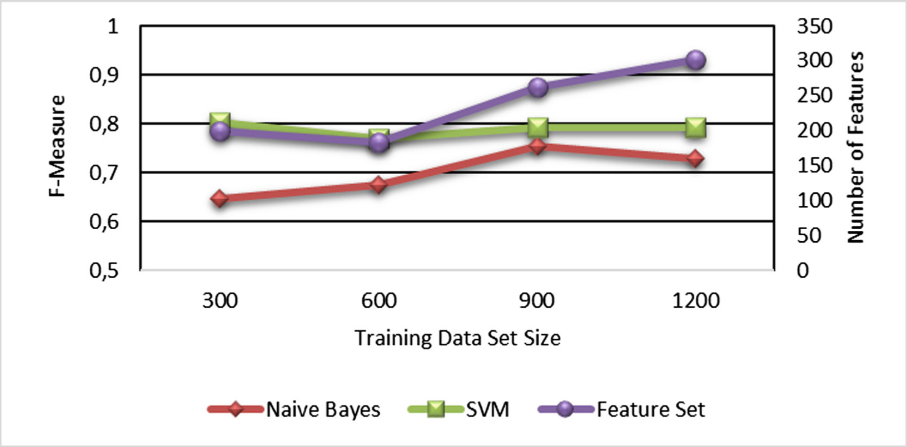

Fig. 2. F-Measure and feature set size over training data set size

Table 1. Performance of sentiment analysis approaches

	Recall	Precision	F-Measure
Naïve Bayes	66.04	81.40	72.92
SVM	71.70	88.37	79.17
SentiStrength	83.02	69.84	75.86

some qualitative assumptions regarding the application effort of the approaches can be made. For *SVM*, a small training data set already seems to allow good analysis accuracy. *Naïve Bayes* needs a larger training data set but still does not match *SVM* performance.

However, there are indicators that the performance of both *SVM* and lexicon based approach can be improved spending effort on data preparation. Looking at the development of the feature set size with regard to the training data set size (Fig. 1), there is a drop from 300 to 600 training data records. This indicates a high diversity in the data. The diversity shows this effect here because of the small sample size. Furthermore, there is an increase in the feature set size from 600 to 1200 training data records, thus no saturation can be seen. The latter again indicates a diversity in the data, assuming more relevant words in an even larger training data set. At last, the extension of the SentiStrength lexicon and the missing effect on accuracy showed, that a diverse vocabulary is used to comment on employers. The selected lexicon elements seem to be semantically relevant but have a low document frequency. The reasons could be a company specific vocabulary or different groups of portal users having different professional backgrounds.

In consequence a larger training data set needs to be used for *SVM* training or a larger amount of comments needs to be screened for relevant lexicon entries in order to have a measurable positive effect on classification accuracy. Regarding the effort of improving the accuracy, manual classification of a text seems to pose less effort than scanning it for lexicon terms, but the effect on accuracy remains unclear at the present state. The use of scales like the star scale by Kununu might reduce the effort of creating training data. Still, the transferability to other data sources must be proven.

Regarding research question 3 “How well do the selected tools/methods evaluate German employer ratings in comparison?”, the *SVM* outperforms *Naïve Bayes* and SentiStrength based on the F-Measure (see Table 1). *SVM* beats *Naïve Bayes* also with regard to Precision and Recall. The picture is not that clear for the pair *SVM* and SentiStrength. If there is a higher weight for the recall because companies do not want to miss negative comments, SentiStrength may be evaluated better. However, application of a 10-fold cross-validation on the training data set resulted in an average recall of 86.33% and an average precision of 81.83% for the *SVM* which is better than the values of SentiStrength. A cross-validation of SentiStrength on the training data set has not yet be performed due to the large effort. Thus, no comparable values exist. Nevertheless, this shows that the performance of *SVM* and SentiStrength could be on nearly the same level. *Naïve Bayes* has been outperformed by *SVM* in all settings (see also Fig. 2).

4 Conclusion and Outlook

This work focused on the possibility of automated sentiment-analysis of German-language comments from the career portal Kununu. Machine learning in the form of *SVM* and lexicon based approaches in the form of SentiStrength showed a good classification performance. There are some indicators that SVM outperforms SentiStrength in the present domain. Nevertheless, following the experimental evaluation in Sect. 3, the accuracy of both methods can be improved by manually preparing data for training or by lexicon construction. However, a cost/benefit estimation seems to be difficult. The rule of thumb by Thelwall et al. in [21] that supposes 2000 records for training data set size is of little help here. Thus, future research should investigate possible guidelines. A way might be provided by the assessment of word frequencies, the use of context information for data selection or by methods for feature selection [31].

Considering the specificities of the domain, a future task will be the evaluation whether the generally shorter length of positive comments is typical for the domain.

Considering the possibility of assessing user provided classification for automated training as in the case of Kununu’s star rating, the application accuracy of the trained model for other sources needs to be assessed. Some bias may be induced by the practise of users to add some positive comments to a negative rating and vice versa.

For future research, an automated content analysis in addition to the sentiment analysis may be interesting. This could help to plan actions based on the evaluation of negative comments. The structure of the comments (headlines/paragraphs) could be included in the analysis here. Furthermore, the general sentiment classification might be improved by using headlines like “Pros” and “Cons”.

In principle, it remains questionable whether the achieved classification quality is sufficient for practical acceptance in the selected application area. There are therefore several directions for future investigations. On the one hand, it is necessary to determine from which classification accuracy is perceived as sufficient (beneficial) for practical use. Then it has to be clarified whether and how the accuracy (possibly by hybrid methods) can be raised to a corresponding level.

References

1. Heyer, G.: Text Mining und Text Mining Services. eDITion, 1/2010:7–10 (2010)
2. Dashtipour, K., Poria, S., Hussain, A., et al.: Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn. Comput. **8**(4), 775 (2016)
3. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
4. Marsland, S.: Machine Learning - An Algorithmic Perspective. CRC Press, Boca Raton (2015)
5. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000). doi:10.1007/3-540-45014-9_1
6. Lämmel, U., Cleve, J.: Künstliche Intelligenz. Carl Hanser Verlag GmbH Co. KG, Munich (2012)
7. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Boston (1999). (Chap. 3)
8. Cherkassky, V., Mulier, F.M.: Learning from Data: Concepts, Theory, and Methods. Wiley, Hoboken (2007)
9. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zero-one loss. Mach. Learn. **29**(2), 103–130 (1997)
10. Rokach, L., Maimon, O.: Top-down induction of decision trees classifiers - a survey. IEEE Trans. Syst. Man Cybern. Part C **35**(4), 476–487 (2005)
11. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl. Based Syst. **89**, 14–46 (2015)
12. Momtazi, S.: Fine-grained German sentiment analysis on social media. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, pp. 1215–1220 (2012)
13. Khan, A., Baharudin, B., Lee, L.H., Khan, K.: A review of machine learning algorithms for text-documents classification. J. Adv. Inf. Technol. **1**(1), 4–20 (2010)
14. Caumanns, J.: A Fast and Simple Stemming Algorithm for German Words. Published in: Department of computer science at the free university of Berlin, pp. 1–10 (1999)
15. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for social web. J. Am. Soc. Inform. Sci. Technol. **63**(1), 163–173 (2012)
16. Blitzer, J., et al.: Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: ACL, pp. 440–447 (2007)
17. Maynard, D., Greenwood, M.A.: Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In: LREC, pp. 4238–4243 (2014)
18. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Found. Trends® Inf. Retr. **2**(1–2), 1–135 (2008). 4.1.2 Subjectivity Detection and Opinion Identification
19. Land, S., Fischer, S.: Rapid miner 5 - rapid miner in academic use (2012). http://docs.rapidminer.com/resources/. Accessed 22 May 2016
20. Shalunts, G., Backfried, G.: SentiSAIL: sentiment analysis in English, German and Russian. In: Perner, P. (ed.) MLDM 2015. LNCS, vol. 9166, pp. 87–97. Springer, Cham (2015). doi:10.1007/978-3-319-21024-7_6

21. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: SentiStrength (2010). http://sentistrength.wlv.ac.uk/. Accessed 16 May 2016
22. Esuli, A., Sebastiani, F.: SENTIWORDNET: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation, pp. 417–422 (2006)
23. Remus, R., Quasthoff, U., Heyer, G.: SentiWS – a publicly available German-language resource for sentiment analysis. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC, pp. 1168–1171 (2010)
24. Cirqueira, D., Jacob, A., Lobato, F., de Santana, A.L., Pinheiro, M.: Performance evaluation of sentiment analysis methods for Brazilian Portuguese. In: Abramowicz, W., Alt, R., Franczyk, B. (eds.) BIS 2016. LNBIP, vol. 263, pp. 245–251. Springer, Cham (2017). doi:10.1007/978-3-319-52464-1_22
25. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D.: Sentiment strength detection in short informal text. J. Am. Soc. Inform. Sci. Technol. **61**(12), 2544–2558 (2010)
26. Hripcsak, G., Rothschild, A.: Agreement, the F-measure, and reliability in information retrieval. J. Am. Med. Inform. Assoc. **12**(3), 296–298 (2005)
27. Balahur, A., Perea-Ortega, J.M.: Sentiment analysis system adaptation for multilingual processing: the case of tweets. Inf. Process. Manag. **51**(4), 547–556 (2015)
28. Kumar, N., Srinathan, K., Varma, V.: Using Wikipedia anchor text and weighted clustering coefficient to enhance the traditional multi-document summarization. In: Gelbukh, A. (ed.) CICLing 2012. LNCS, vol. 7182, pp. 390–401. Springer, Heidelberg (2012). doi:10.1007/978-3-642-28601-8_33
29. Scharkow, M.: Thematic content analysis using supervised machine learning: an empirical evaluation using German online news. Qual. Quant. **47**(2), 761–773 (2011)
30. Scholz, T., Conrad, S., Wolters, I.: Comparing different methods for opinion mining in newspaper articles. In: Bouma, G., Ittoo, A., Métais, E., Wortmann, H. (eds.) NLDB 2012. LNCS, vol. 7337, pp. 259–264. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31178-9_31
31. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowl. Inf. Syst. **34**(3), 483–519 (2013). doi:10.1007/s10115-012-0487-8