Sentiment Analysis of German Emails: A Comparison of Two Approaches

Bernd Markscheffel and Markus Haberzettl
Department of Information and Knowledge Management, Technische Universität Ilmenau, Ilmenau, Germany

Keywords: Sentiment Analysis, Literature Analysis, Machine Learning, Feature Extraction Methods.

Abstract: The increasing number of emails sent daily to the customer service of companies confronts them with new challenges. In particular, a lack of resources to deal with critical concerns, such as complaints, poses a threat to customer relations and the public perception of companies. Therefore, it is necessary to prioritize these concerns in order to avoid negative effects. Sentiment analysis, i.e. the automated recognition of the mood in texts, makes such prioritisation possible. The sentiment analysis of German-language emails is still an open research problem. Moreover, there is no evidence of a dominant approach in this context. Therefore two approaches are compared, which are applicable in the context of the problem definition mentioned. The first approach is based on the combination of sentiment lexicons and machine learning methods. This is to be extended by the second approach in such a way that in addition to the lexicons further features are used. These features are to be generated by the use of feature extraction methods. The methods used in both approaches are investigated in a systematic literature search. A Gold Standard corpus is generated as basis for the comparison of these approaches. Systematic experiments are carried out in which the different method combinations for the approaches are examined. The results of the experiments show that the combination of feature extracting methods with Sentiment lexicons and machine learning approaches generates the best classification results.

1 INTRODUCTION

One of the preferred communication channels in the field of customer service is email (Gupta et al., 2010). The increasing number of emails arriving daily at customer service, therefore, poses a challenge for the prompt processing of customer concerns in companies (Radicati Group, 2018) Automated prioritization is necessary in order to identify and prioritize critical concerns to avoid the risk of negative effects on the perception of companies.

One form of prioritization is the sentiment, the emotionally annotated mood and opinion in an email (Borele and Borikar; 2016). A sentiment is also an approach to solving further problems such as the analysis of the course of customer contacts, email marketing or the identification of critical topics (Nasukawa and Yi, 2003). Linguistic data processing (LDV) approaches are used to automatically capture sentiment (Agarwal et al., 2011).

Although the number of published research papers is increasing, sentiment analysis continues to be an open research problem (Bravo-Marquez, Mendoza and Poblete, 2014; Ravi and Ravi, 2014), in particular, there is a lack of in approaches specifically for the German language, whereby the automated classification of polarity in the categories positive, negative and neutral is of particular interest (Scholz et al., 2012; Steinbauer and Kröll, 2016; Waltinger, 2010). In research, methods of machine learning have prevailed over knowledge- and dictionary-based methods to determine polarity (Scholz et al., 2012). The reason for this is that machine learning methods approach human accuracy and are not restricted by the other two approaches (e.g. lack of dynamics in relation to informal language) (Cao et al., 2015; Sebastiani, 2002). In contrast to knowledge- and dictionary-based methods, which are manual rule definitions, machine learning represents the fully automated inductive detection of such rules using algorithms developed for this purpose (Sebastiani, 2002). So far, no machine learning method or procedures and approaches based on it have been identified as dominant - another reason why sentiment analysis is today still an unsolved research problem (Vinodhini and Chandrasekaran, 2012; Argamon et al., 2007; Borele and Borikar, 2016).

Markscheffel, B. and Haberzettl, M.
Sentiment Analysis of German Emails: A Comparison of Two Approaches.
DOI: 10.5220/0008114803850391
In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*, pages 385-391
ISBN: 978-989-758-377-3

One solution for the classification of polarity is seen (A) in the combination of sentiment dictionaries and machine learning methods (Ohana and Tierney, 2009). Further potential is considered (B) in the combination of such lexicons and learning methods with other methods of feature extraction (Ohana and Tierney, 2009).

The main aim of this paper is to compare these two approaches for German-language emails at the document level to answer the question, do machine learning methods based on sentiment lexicons (A) generate better results in the context of sentiment analysis if the lexicon is combined with other methods of feature extraction (B).

2 WORKFLOW

The several machine learning and feature extraction methods to be identified for the different approaches are determined by a systematic literature analysis according to Webster and Watson (2004) and is additionally supplemented by Prabowo and Thelwall (2009) when structuring the findings. The complete results of the literature analysis, the determined machine learning methods, and the identified relevant feature extraction methods can be found in Haberzettl and Markscheffel (2018). The implementation of these approaches to be compared is done with the Konstanz Information Miner (KNIME) in version 3.5.2.25. The data required for implementation are acquired according to the Gold Standard requirements of Wissler et al., (2014). The results of the approaches will then be compared using identified quality criteria which have been recognized in the context.

2.1 Data Aquisition

Because text data, i.e. unstructured data, is to be classified in sentiment analysis, it must be converted into structured data for the real classification process. This data is collected in a corpus and split into a training data set and a test data set for the analysis process. In the absence of a suitable freely accessible corpus for this task, a separate corpus has to be acquired and coded which fulfills Gold Standard requirements.

For this purpose, 7,000 requests from private customers to the customer service of a company in the telecommunication sector are used. Since a full survey is not possible due to the manual coding effort and no information on the distribution of polarity in the population is available, this sample was determined based on a simple random selection. Coding by only one expert should be rejected, especially in view of the Gold Standard requirement. The argumentation for a higher data consistency due to this is to be critically considered especially in light of the subjectivity of the sentiment - sentiment is interpreted differently by different persons, for example, due to different life experiences (Nasukawa and Yi, 2003; Bütow et al., 2017; Calzolari et al., 2012; Thelwall et al., 2010). This characteristic has to be reflected in the corpus. The following parameters, therefore, apply to the coding: Emails should be evaluated from the writer's point of view and categorized exclusively as an entire document. In addition, only subjective statements are relevant for determining positive or negative sentiments. The coding was therefore carried out in three phases:

In the first phase, the sample was divided into seven equally sized data sets. These groups were coded by six different experts who had previously received a codebook with instructions (the assignment of the groups was random in each phase, however, no reviewer coded a document twice). In addition to the general conditions, the codebook contains the class scale to be used and instructions for the classification of the classes:

- 1 (very positive)
- 2 (positive)
- 3 (neutral)
- 4 (negative)
- 5 (very negative)
- Mixed (contains positive and negative elements).

Due to the subjective interpretation of the sentiment, the groups were again coded by different experts in a second step. This expert had no information about the previous coding.

In phase three, all emails were identified, which were coded differently in each of the previous phases. These emails were assigned to a new expert for the group, who performed a third encoding.

The corpus is then divided into a training and test data set in a stratified manner with a ratio of 70:30. The emails are then converted into documents.

2.2 Data Preprocessing

In the source system, the emails are already pre-processed: Personal customer data (name, address, etc.) have been anonymized and replaced. HTML tags, meta data (sender, IDs, etc.), attachments have been deleted and message histories in the emails removed. Nevertheless, there is a large number of non-text elements to be found, which therefore have to be eliminated.

The pre-processing workflow consists of ten steps.

1. Word separation: unintentionally moved words must be separated - an error that occurs during database loading.
2. Replace umlauts: ä → ae, ö → oe, ü → ue; ß → ss.
3. Dictionary-based lemmatization: the transformation of inflected words back into their basic form, freust → freuen.
4. Text normalization via lower casing.
5. Named entity recognition: iPhone 6 Plus → iphone.
6. Character limitation: only characters (a, b, c…) are allowed.
7. Spelling error correction with the help of the Wiktionary Spelling Error Dictionary.
8. Stop word elimination.
9. Removal of word <= 3 characters.
10. Output is the pre-processed, tokenized corpus, ready for the comparison tasks.

2.3 Feature Extraction and Selection

The next step is to extract features from this corpus. Features are defined as numerically measurable attributes and properties of data. In the context of text mining, feature extraction should, therefore, be understood as the structuring process of unstructured data; the methods are used to identify and extract structured data in unstructured data. The extraction is split into two parts: Features are generated on the one hand by direct conversion of texts or tokens and on the other hand by applying the feature extracting methods identified and introduced in Haberzettl and Markscheffel (2018). Table 1 illustrates the several feature extraction methods used in our approach.

Table1: Feature extracting methods.

n-Gramm (n-G)	Negation (Neg)
Term frequency - Inverse document frequency (TF-IDF)	Pointwise Mutual Information (PMI)
Term presence (TP)	Sentiment Dictionary (SM)
Term frequency (TF)	Category (Cat)
Part of speech tagging (POS)	Corpus specific
Modification feature (MF)	

The conversion takes place in text mining usually on the basis of the Bag-Of-Words (BoW). After the conversion, no more documents exist accordingly, (the structured data were "extracted" from the documents in a sensual way). Instead, the documents are represented by a document vector. The document vector contains the feature vector, i.e. the vector of all extracted features.

2.4 Sentiment Lexicon

Sentiment dictionaries are required as a basis for the approaches A and B described above. Sentiment dictionaries are dictionaries in which words are assigned to a polarity index. Sentiment dictionaries are context-sensitive, i.e. words and values contained in them apply primarily to the context in which they were created. Since no suitable dictionary exists for the context of German-language emails, such a dictionary had to be created. For resource reasons, an automated, corpus-based approach was pursued. According to SentiWS (Remus et al., 2010), a generation on coocurrency based rule is chosen. Pointwise Mutual Information (PMI) is used as a method for the analysis of coocurrency and thus for the determination of semantic orientation (Remus et al., 2010; Turney, 2002; Turney and Littmann, 2003). In our specific case, two million uncoded emails were acquired from the same database as the corpus. The selection was made by random sampling. All emails were pre-processed according to the process described above.

Table 2: Cut-out of the Sentiment Dictionary SentiMail (SM).

Positive Term	Scaled PMI	Negative Term	Scaled PMI
herzlich	1	betruegen	-1
empathisch	0,9786	verarschen	-0,983
begleueck-wuensche	0,9589	andrehen	-0,9798
angenehm	0,954	dermassen	-0,9743
bedanken	0,9259	vertrauens-bruch	-0,9628
kompliment	0,9156	scheiss	-0,9336
danke	0,9148	anluegen	-0,9263
sympathy-schen	0,9134	abzocke	-0,9233
sympathisch	0,8956	taeuschung	-0,9181
nervositaet	0,878	geschaefts-gebaren	-0,9137

For all words contained in these emails the semantic orientation {negative, positive} was determined on the basis of the PMI (Remus et al., 2010; Turney, 2002), i.e. for each word its similarity to previously defined positive or negative seed words is calculated. For each of the 93,170 words identified, a threshold value for clipping the lexicon SO-PMI ∈ [-0,13;0,08] was determined by manual checking, taking into account the Zipf distribution, so that the final lexicon consists of 955 positive and 1,704 negative words. Table 2 shows a cut-out of the sentiment dictionary with its top ten positive and negative normalized PMI-values, whereby the normalization is within the boundaries of PMI ∈ [-1;1].

3 EXPERIMENTS AND RESULTS

For the implementation of the machine learning methods to be investigated (Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes (NB), Logistic Regression (LR) or Maximum Entropy (ME) and k-nN nearest neighbour (k-nN) cf. Haberzettl and Markscheffel (2018) in combination with the above mentioned feature extracting methods different libraries of Weka integration of KNIME were used (e.g. LibSVM; NaiveBayesMultinominal) or could be used directly as nodes (LR Logistics (3,7), k-nN). The ANN was implemented by a multi-layer perceptron starting from our multi-class case. One layer and M/2 (M=feature) neurons in this layer were chosen as a starting point and then successively increased to M+2 neurons.

3.1 Evaluation

The results of the experiments and thus the classification itself are to be evaluated with the use of quality criteria. With the help of a confusion matrix, the results of the classification can be divided according to positive and negative cases. The four resulting cases from the classification in the confusion matrix (true positive, true negative, false positive, false negative) allow the derivation of the following different quality criteria: Accuracy (ACC), Precision (PRE), Recall (REC) and F-Measure (F1) (Cleve and Lämmel, 2014, Davis and Goadrich, 2006). The validity of the quality criteria is ensured by a 10-fold stratified cross-validation (Kohavi, 1995). Accuracy is used as the decisive criterion for determining the best result due to the limitations discussed in Haberzettl and Markscheffel, (2018).

3.2 Experiments and Results for the Sentiment Dictionary (A) and Feature Extraction (B)

In a first step, based on the approaches A and B, the sentiment lexicon to be used was first determined. For this purpose, all learning methods were trained on the features of SentiWS, SentiMail and the combination of both. The result is the result of assumption A. Figure 1 shows the corresponding workflow implemented with KNIME for experiments A and B.

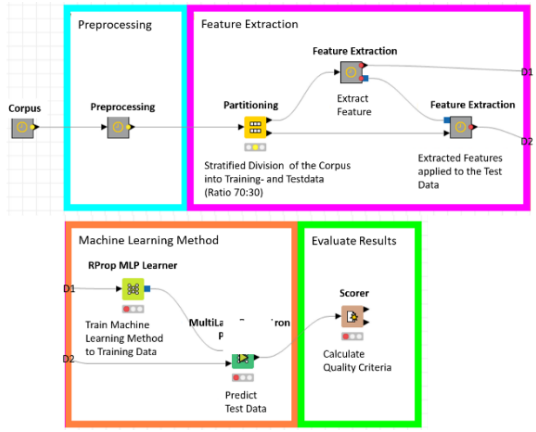

Figure 1: KNIME Workflow for the Experiments A and B.

The results of the first step are obvious: For each learning method, the combination of both sentiment lexicons is the best alternative with regard to each quality criterion. Only the precision at NB is better with SentiWS - probably, measured by the recall, due to the simple assignment of the emails to the most frequented class (neutral).

Table 3: Comparison of the sentiment lexicons SentiMail (SM) and SentiWS (SW) as a feature extraction method and the best result (rank (R), evaluated according to Accuracy) for approach A.

	R	ACC	PRE	REC	F1	
SVM	2	83,19%	83,26%	71,43%	75,87%	SM SW
	5	80,41%	80,18%	63,76%	69,14%	SM
	9	78,44%	74,70%	62,25%	65,86%	SM
ANN	1	83,82%	82,26%	74,55%	77,78%	SM SW
	4	81,44%	79,68%	68,20%	72,49%	SM
	7	79,17%	73,64%	65,98%	68,83%	SM
NB	12	75,67%	68,47%	67,81%	67,16%	SM SW
	14	74,47%	67,45%	63,95%	63,92%	SM
	15	72,89%	71,59%	47,41%	49,75%	SM
ME	3	82,73%	82,22%	70,97%	75,21%	SM SW
	6	80,33%	79,35%	64,22%	69,32%	SM
	10	78,32%	74,30%	61,74%	65,72%	SM
KnN	8	79,14%	75,01%	69,23%	71,65%	SM SW
	11	77,58%	71,81%	65,75%	68,22%	SM
	13	75,09%	67,88%	61,70%	64,08%	SM

Particularly, with regard to the exactness (Precision, Recall, F1-Measure), the combination of both lexicon is dominant. Table 3 shows a compilation of the results.

So, out of the results of experiment A both sentiment lexicon were selected from the results of A. It should be noted that the SentiMail (SM) lexicon, created within the context, produces better results in direct comparison with SentiWS (SW) - this substantiates the need for context-dependent sentiment dictionaries. The rank assigned according to Accuracy indicates that the best result for experiment A is the combination of ANN and both sentiment dictionaries. This result is also confirmed by the remaining quality criteria (F1 is to be weighted higher than the Precision outlier).

Table 4: Comparison of term presence (TP) vs. TF-IDF vs. relative term frequency (relTF) as additional features to A.

	R	ACC	PRE	REC	F1	
SVM	1	84,67%	80,93%	76,65%	78,59%	TP
	2	84,16%	84,38%	73,48%	77,73%	TF-IDF
	3	83,73%	83,99%	72,55%	76,93%	Rel TF
ANN	7	77,02%	67,17%	67,01%	67,06%	TP
	8	76,92%	67,40%	65,98%	66,65%	TF-IDF
	9	75,72%	65,48%	66,39%	65,91%	Rel TF
NB	4	81,32%	74,08%	78,26%	75,96%	TP
	5	78,83%	71,86%	72,95%	72,23%	TF-IDF
	6	77,87%	71,15%	70,51%	70,56%	Rel TF
ME	10	72,83%	61,79%	67,14%	63,75%	TP
	12	71,33%	59,98%	64,51%	61,77%	TF-IDF
	13	71,30%	60,04%	64,86%	61,91%	Rel TF
KnN	11	72,43%	70,51%	52,80%	51,34%	TP
	14	69,29%	58,60%	59,01%	54,93%	Rel TF
	15	68,28%	56,88%	60,59%	55,05%	TF-IDF

For the second experiment (B), the best lexicon for each learning method is used. The next step is to determine which frequency is to be used for the unigrams. The background for this is the frequently cited comparison between term presence (TP) and relative term frequency (relTF), at which the term presence dominates (Pang and Lee, 2008). For this purpose, each machine learning method was trained with all three frequency types (TP, relTF, TF-IDF) in each case as well as the identified sentiment lexicons from the previous experiment step. For the next step, only the frequency with which each learning method achieves the best results according to Accuracy was selected for each learning method. The results of the remaining 62 possible combinations of the feature categories for each learning method are evaluated, whereby each of these combinations must inevitably contain the sentiment dictionary and produces the results for experiment B.

How to recognize from Table 4 the values for term presence (TP) are better than the values for TF-IDF as well as to the relative term frequency (relTF). Accordingly, in the next step, only the term presence for unigrams was used for all machine learning methods. At this point, the results that significantly vary from the previous stage should be highlighted. Thus, the accuracy of the previously best learning method (ANN) decreases by 6.8 percentage points, while, for example, the accuracy of the SVM (F1-Measure) increases further. This mainly reflects the core characteristics of the SVM, which benefits significantly more from large feature vectors than other learning methods. Also noteworthy is the small difference between TF-IDF and relTF. Although four of the five learning methods achieved a higher accuracy with TF-IDF than with the relative term frequency, the results of the quality criteria between the two frequencies usually deviate only marginally. As Table 5 shows, the results of SVM as well as of NB and ME with approach B are significantly better with regard to Accuracy and F1-Measure than in approach A. In particular, the 6.6 percentage points higher accuracy and the 9.78 percentage points higher F1 measurement at NB should be highlighted. ANN and k-nN show no significant deviations from A, whereby the ANN generates marginally worse results with respect to almost all quality criteria than in approach A.

Table 5: Best results for experiment B (rank, measured by Accuracy), i.e. for features in combination with SentiWS and SentiMail.

	R	ACC	PRE	REC	F1
SVM	1	85,03%	81,22%	77,98%	79,49%
POS, Neg, n-G					
ANN	2	83,64%	81,84%	74,79%	77,83%
TF					
NB	4	82,27%	75,62%	78,44%	76,94%
POS, Booster, Neg, n-G					
ME	3	83,28%	81,43%	72,52%	76,14%
TF, POS, Cat					
KnN	5	79,95%	77,22%	68,26%	71,77%
TF					

4 SUMMARY AND FUTURE WORK

On the background of optimizing the analysis of the polarity of German-language emails at the document level, two approaches to sentiment analysis were compared in experiments: Approach A combines machine learning methods and sentiment dictionaries. Approach B extends this with additional feature extraction methods. Measured against the quality criteria of the best results per approach, approach B dominates in three of four cases (exception precision) over A (see Table 6).

Table 6: Comparison of the best results of approach A and B.

		ACC	PRE	REC	F1	
A	ANN	83,82%	82,26%	74,55%	77,78%	SWSM
B	SVM	85,03%	81,22%	77,98%	79,49%	SWSM
	corresponding feature extraction method			POS, Neg, n-G		

When analyzing the results of the individual experiments, a dependence of the results on the selected feature extraction and machine learning methods or feature combinations can be noticed. In a further approach, it can be explored to what extent multi-layered methods of supervised or unsupervised machine learning can improve the results. At least according to Stojanowski (2015), the automation of feature extraction makes deep learning in the context of sentiment analysis more flexible and robust than classical approaches when applied to different domains (language, text structure, etc.).

This approach allows for further improvements. We have also implemented this approach and, as expected, it generated even better results than the hybrid approaches presented here. A detailed description of the methodology used and the results will be the subject of further work.

REFERENCES

Agarwal, A. et al., 2011. Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Language in Social Media LSM 2011*. Portland, pp. 30-38.

Argamon, S. et al., 2007. Automatically Determining Attitude Type and Force for Sentiment Analysis. In Z. Vetulani and H. Uszkoreit, (eds.). *Human Language Technology. Challenges of the Information Society. Third Language and Technology Conference, LTC 2007*. LNAI. Poznan, pp. 218-231.

Borele, P., Borikar, D., 2016. An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques. *IOSR Journal of Computer Engineering*, 18(2), pp. 64-69.

Bravo-Marquez, F., Mendoza, M., Poblete, B., 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69, pp. 86-99.

Bütow, F., Schultze, F., Strauch, L., 2017. *Semantic Search: Sentiment Analysis with Machine Learning Algorithms on German News Articles.* Available from: http://www.dai-labor.de/fileadmin/Files/Publikationen/Buchdatei/BuetowEtAl--SentimentAnalysisOnGermanNews.pdf. [Accessed: 01/02/2019].

Calzolari, N. et al. (eds.), 2012. *Proceedings of the 8th International Conference on Language Ressources and Evaluation (LREC-2012)*. Istanbul, pp. 1215-1220.

Cao, Y., Xu, R., Chen, T., 2015. Combining Convolutional Neural Network and Support Vector Machine for Sentiment Classification. In X. Zhang et al. (eds.). *Social Media Processing. SMP 2015. Communications in Computer and Information Science*, 568, Guangzhou, pp. 144-155.

Cleve, J., Lämmel, U., 2014. *Data Mining*. München, 2014.

Davis, J., Goadrich, M., 2006. The Relationship Between Precision-Recall and ROC Curves. In W. Cohen, A. Moore (eds.). *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, pp. 233-240.

Gupta, N., Gilbert, M., Di Fabbrizio, G., 2010. Emotion Detection in Email Customer Care. In D. Inkpen, C. Strapparava (eds.). *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, pp. 10-16.

Haberzettl, M., Markscheffel, B., 2018. A Literature Analysis for the Identification of Machine Learning and Feature Extraction Methods for Sentiment Analysis. In *Proceedings of the 13th International Conference on Digital Information Management (ICDIM 2018)*, Berlin.

Kohavi, P., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In C. S. Mellish (eds.). *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*. Montreal, pp. 1137-1143.

Nasukawa, T., Yi, J., 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In Gennari, J., Porter, P., Gil, Y., (eds.). *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP'03).* Sanibel Island, pp. 70-77.

Ohana, B., Tierney, B., 2009. Sentiment Classification of Reviews Using SentiWordNet. In *9th IT&T Conference*, Dublin, pp. 1-9.

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval*. 2(1-2), pp. 1-135.

Prabowo, R., Thelwall, M., 2009. Sentiment analysis: A combined approach. *Journal of Informetrics.* 3, pp. 143-157.

Radicati Group, 2018. *Email Statistics Report, 2018-2022 – Executive Summary*. Available from https://www.radicati.com/wp/wp-content/uploads/2018/01/Email_Statistics_Report,_2018-2022_Executive_Summary.pdf, 2018, [Accessed: 01/04/2018].

Ravi, K., Ravi, V., 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89, pp. 14-46.

Remus, R., Quasthoff, U., Heye, G., 2010. SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. In *International Conference on Language Resources and Evaluation*. pp. 1168-1171.

Scholz, T., Conrad, S., Hillekamps, L., 2012. Opinion Mining on a German Corpus of a Media Response Analysis. In P. Sojka, et al. (eds.). *Text, Speech and Dialogue, 15th International Conference, TSD 2012*. Brno, pp. 39-46.

Sebastiani, F., 2002. Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)*, 34(1), pp. 1-47.

Steinbauer, F., Kröll, M., 2016. Sentiment Analysis for German Facebook Pages. In E. Métaiset et al. (eds.). *Natural Language Processing and Information Systems, 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016*. Salford, pp. 427-432.

Stojanovski, D. et al., 2015. Twitter Sentiment Analysis Using Deep Convolutional Neural Network. In E. Onieva et al. (eds.). *Hybrid Artificial Intelligent Systems. HAIS 2015. Lecture Notes in Computer Science. 9121*, pp. 726-737.

Thelwall, K. et al., 2010. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*. 61(12), pp. 2544-2558.

Turney, P.D., 2002. Thumbs Up oder Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In P. Isabelle (eds.). *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, pp. 417-424.

Turney, P.D., Littman, M.L., 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*. 21(4), pp. 315-346.

Vinodhini, G., Chandrasekaran, R.M., 2012. Sentiment Analysis and Opinion Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering. 2(6) June, pp. 282-292.

Waltinger, U., 2010. GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis. In N. Calzolari, et al. (eds.). *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC-10)*, Valletta, pp. 1638-1642.

Webster, J., Watson, R.T., 2002. Analyzing the Past to Prepare for the Future: Writing a Literatur Review. *MIS Quarterly*. 26(2), pp. 13-23.

Wissler, L. et al., 2014. The Gold Standard in Corpus Annotation. In *Proceedings of the 5th IEEE Germany Student Conference. Institute of Electrical and Electronics Engineers - IEEE*. Passau, pp. 1-4.