Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1627–1632
Marseille, 11–16 May 2020

Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, Hans-Joachim Böhme
University of Applied Science (HTW) Dresden, Germany
T2K: Text to Knowledge, Dresden, Germany
{oliver.guhr, frank.bahrmann, hans-joachim.boehme}@htw-dresden.de
anne-kathrin.schumann@text2knowledge.de

Abstract

This paper describes the training of a general-purpose German sentiment classification model. Sentiment classification is an important aspect of general text analytics. Furthermore, it plays a vital role in dialogue systems and voice interfaces that depend on the ability of the system to pick up and understand emotional signals from user utterances. The presented study outlines how we have collected a new German sentiment corpus and then combined this corpus with existing resources to train a broad-coverage German sentiment model. The resulting data set contains 5.4 million labelled samples. We have used the data to train both, a simple convolutional and a transformer-based classification model and compared the results achieved on various training configurations. The model and the data set will be published along with this paper.

Keywords: sentiment analysis, German sentiment model, German corpus

1. Introduction

The presented work is carried out within the context of developing service robots for public spaces and nursing homes. The robots are deployed in different scenarios, for instance as tour guides in museums and as assistants in health care settings. In both situations, it is crucial that the robot, or more precisely, its voice user interface, is able to adapt to users' needs and intelligently respond to users' emotions. We believe that the ability to pick up and understand emotional signals provided by the users' language input is crucial not only for the acceptance of service robots but also for the handling of situations where the users provide strong emotional feedback.

There are two main applications for sentiment analysis in the context of a robots' conversational voice interface:

- We want to classify the sentiment of all user statements. This will enable the dialogue manager of the robot to respond to positive and negative user feedback.
- Moreover, the users' utterances and their sentiment labels can later be used to rate conversations and evaluate the users' reactions to changes in the behaviour of the robot. This concept is common and other conversational voice interfaces like (Chen et al., 2018) and (Fang et al., 2018) also apply sentiment analysis for these tasks.

The development of reliable machine learning models for the mentioned tasks requires annotated training data. However, available German sentiment data sets such as *PotTS* (Sidarenka, 2016), *SB10k* (Cieliebak et al., 2017), and *GermEval-2017* (Wojatzki et al., 2017), when combined, comprise only 39,000 sentences.

In contrast, commonly used English data sets such as *IMDB* (Maas et al., 2011) contain 50,000 binary labeled items and the *YELP* data set (Zhang et al., 2016) contains 598,000 binary labeled items. Another aspect that is worth noticing is that both *PotTS* and *SB10k* consist of annotated tweets, and *GermEval-2017* contains a combination of Tweets and Facebook posts. Therefore, all available German data sets reflect social media language rather than general-purpose German.

To overcome these limitations, we have collected additional data by crawling hotel reviews from the popular German website *holidaycheck.de* and by crawling movie reviews from the website *filmstarts.de*. The ratings contained in these reviews were adapted to fit with the three sentiment classes (negative, neutral, positive) of the already existing data sets. We applied the same process to the *Scare* Corpus (Sänger et al., 2016) which consists of mobile app ratings. Additionally, we enhanced the *neutral* class by adding texts from the *Leipzig corpora* collection (Goldhahn et al., 2012). In the last step, we further enhanced the data set by adding utterances that contain strongly emotional vocabulary such as insults. These utterances have been recorded in the course of field experiments with the service robots (Poschmann et al., 2012; Hellbach et al., 2013).

Afterwards, we combined all the data sets to create a large, and broad-coverage, German sentiment data set that is a better approximation of the types of utterances that can be directed to service robots. We then trained two different types of sentiment classification models, namely FastText (Joulin et al., 2016) and BERT (Devlin et al., 2019) and evaluated there performance. The data sets, trained models and the source code are publicly available [1].

2. Related Work

There are several sentiment dictionaries for German, for instance BAWL-R, SentiWS and GermanPolarityClues (Võ et al., 2009; Remus et al., 2010; Waltinger, 2010). Existing corpora cover social media language (*PotTS*, *SB10k*, *GermEval-2017*) and app reviews (*Scare*).

[1] https://github.com/oliverguhr/german-sentiment

Cieliebak et al. applied SVM- and CNN-based machine learning models to their *SB10k* data set. They have achieved a maximum F_1 score (average over positive and negative F_1 scores) of 65.09 on their data set.
A recent initiative in sentiment classification for German is the *GermEval-2017* shared task on aspect-based sentiment classification (Wojatzki et al., 2017). It attracted more than 50 submissions. The data set contains utterances that refer to the German train operator "Deutsche Bahn". The task offered several sub-tasks, such as relevance filtering, document-level and aspect-level sentiment classification as well as opinion target extraction.
For document-level sentiment classification, the best performing system reached a micro-averaged F_1 score of 74.9. This approach (Naderalvojoud et al., 2017) is particularly interesting because it incorporates information from existing sentiment lexica into a neural network architecture.
Schmitt et al. (2018) published the *GermEval-2017* data set. They have experimented with both Bi-LSTMs and CNNs to carry out end-to-end aspect-based sentiment analysis. The goal of their work was to predict the sentiment of an input sentence with respect to a finite set of aspects of interest. They have achieved a maximum micro-average F_1 score of 46.5 (which is significantly above the best GermEval result) for GermEval's joint aspect and sentiment task, using FastText embeddings and an end-to-end CNN architecture.

3. Data Set

We chose to use three classes: positive, neutral and negative for our data set. The neutral class is important because not all of the users' utterances are expected to contain a positive or negative sentiment. Questions like "What is your name?" or "Can you help me?" do not contain any sentiment and should be classified as neutral. The three manually annotated data sets *PotTS*, *SB10k* and *GermEval-2017* where also labeled with this classes. To expand the variety of domains and improve the overall quantity of our data set we have replicated the approach described by (Pang et al., 2002). We obtained reviews of hotels, movies, and apps from sources, that allow their users to add a star rating to their reviews. These ratings are in the range of zero to five stars, where zero or one-star denotes the most negative and five stars the most positive review. Like Pang, we chose to use only positive and negative reviews for our data set. We considered reviews with less than 3 stars as negative and reviews with more than three stars as positive, all reviews with three stars were omitted. To fill the gap of neutral training samples, we used texts from *Leipzig corpora* collection. This data set consists of crawled news texts and Wikipedia articles.

3.1. Data Sources

PotTS (Sidarenka, 2016) contains 7,504 messages from the social media platform Twitter. The messages were collected in 2013 and manually labelled by two experts. The authors used a keyword filter to select tweets from the following topics: federal elections in Germany in 2013, papal conclave in 2013, discussions about general political issues and casual everyday conversations.

SB10k (Cieliebak et al., 2017) contains 9,783 labeled German tweets. The tweets were collected in 2013 between August and October. Each tweet was labelled by 3 human annotators using five classes: positive, negative, neutral, mixed, and unknown. However, the authors published not the full-text of the tweets, but the IDs of the annotated tweets. When we downloaded the tweets in 2018 from Twitter using the IDs, a substantial portion of the tweets were no longer accessible. Therefore we decided to use an earlier collected version[2] of the data set, containing 7,474 full-text tweets. For our experiments, we kept only the positive, negative, and neutral classes.

GermEval-2017 (Wojatzki et al., 2017) was published as a part of the *GermEval-2017* shared task on aspect-based sentiment analysis. The data set consists of documents from social media, microblogs, news, and Q&A sites about the German train operator "Deutsche Bahn". Every document was labelled by two trained annotators. Overall the data set contains almost 28,000 documents, from which 23,525 documents are available for public use. The documents have been collected between May 2015 and June 2016.

Scare (Sänger et al., 2016), the Sentiment Corpus of App Reviews, contains over 800,000 application reviews from the Google Play Store. The corpus consists of reviews from 148 applications across eleven categories (a detailed list can be found at the corpus website.[3]). The authors stated, that these reviews are shorter than typical product reviews and that they use colloquial language and a more flexible grammar. All reviews were retrieved between December 2014 and June 2015.

The **Filmstarts** data set consists of 71,229 user written movie reviews in the German language. We have collected this data from the German website *filmstarts.de* using a web crawler. The users can label their reviews in the range of 0.5 to 5 stars. With 40,049 documents the majority of the reviews in this data set are positive and only 15,610 reviews are negative. All data was downloaded between the 15th and 16th of October 2018, containing reviews up to this date.

The **holidaycheck** data set contains hotel reviews from the German website *holidaycheck.de*. The users of this website can write a general review and rate their hotel. Additionally, they can review and rate six specific aspects: location & surroundings, rooms, service, cuisine, sports & entertainment and hotel. A full review contains therefore seven texts and the associated star rating in the range from zero to six stars. In total, we have downloaded 4,832,001 text-rating pairs for hotels from ten destinations: Egypt, Bulgaria, China, Greece, India, Majorca, Mexico, Tenerife, Thailand and Tunisia. The reviews were obtained from November to December 2018 and contain reviews up to

[2] https://github.com/WladimirSidorenko/CGSA
[3] http://www.romanklinger.de/scare

Data Set	Positive Samples	Neutral Samples	Negative Samples	Total Samples
Emotions	188	28	1,090	1,306
Filmstarts	40,049	0	15,610	55,659
GermEval-2017	1,371	16,309	5,845	23,525
holidaycheck	3,135,449	0	388,744	3,524,193
leipzig-wikipedia	0	1,000,000	0	1,000,000
PotTS	3,448	2,487	1,569	7,504
SB10k	1,716	4,628	1,130	7,474
Scare	538,103	0	197,279	735,382
Sum	**3,720,324**	**1,023,452**	**611,267**	**5,355,043**

Table 1: This table shows the list of data sets that we used and the number of samples per class. Note that these are numbers relate to the data sets after the preprocessing.

this date. After removing all reviews with no stars or four stars, the data set contains 3,524,193 text-rating pairs.

leipzig-wikipedia was taken from a corpus collection published by (Goldhahn et al., 2012) and was used to fill the gap of neutral labelled texts. This data set contains sentences from Wikipedia. We argue that these texts are in a neutral tone and therefore do not contain any sentiment. This data set contains 1,000,000 documents, which we have labelled as neutral. For this work, we have used the latest version from 2016.

The **Emotions** data set contains a list of utterances that we have recorded during the "Wizard of Oz" experiments with the service robots. We have noticed, that people used insults while talking to the robot. Since most of these words are filtered in social media and review platforms the other data sets do not contain such words. We used synonym replacement as a data augmentation technique to generate new utterances based on our recordings. Besides negative feedback, this data set contains also positive feedback and phrases about sexual identity and orientation that where labelled as neutral. Overall this data set contains 1,306 examples.

3.2. Data processing

We have applied a three-step process to create the final data set. Every source data set was preprocessed, splitted into classes and then recombined.

The **preprocessing** comprises following steps: We removed all URLs and user names starting with a @ character and replaced all numbers with numerals. Furthermore, we removed all non-German characters and punctuation characters. We also replaced smileys and emoticons with sentiment tags. For this replacement, we have used the dictionaries that were provided as part of the corpus by (Sänger et al., 2016). Samples that did not contain any characters after the preprocessing were removed. Table 1 shows the number of samples per class per data set after this step.

We choose to **split the data** into a training, validation and test set. We used 70% of the data for training the model, 20% for hyperparameter optimisation on the validation set and 10% of the data for the test set to compare the different models. To maintain the different portions of the three classes, we applied this three-way splitting for every class.

We created two versions of the data set, one unbalanced data set containing all 5.355 million samples and one balanced data set containing 1.834 million samples. We created a balanced data set using downsampling, it contains 611,267 samples in every class.

4. Sentiment Classification Models

To train sentiment classification models, we chose two different machine learning approaches. We evaluated the performance of FastText and BERT using the micro averaged F_1 score and confusion matrices. Moreover, we have trained both models on the balanced and unbalanced version of our data set to test if the model becomes biased towards classes with more training samples. To compare the models' performance on the individual source data sets, we kept the name of the source data set for every sample in the test set. This way we were able to compute the F_1 scores for the individual source sets.

Since the F_1 score can only be used for binary classification tasks, we use the micro and macro averaged F_1 score as defined in (Sebastiani, 2002, p. 33). The micro F_1 score is a weighted average of the F_1 score of all three classes. The macro F_1 score is the average of each class F_1 score.

4.1. FastText

FastText is an improvement over traditional word embeddings. In that it is based on a bag of n-grams rather than a bag of words (Bojanowski et al., 2017): This allows FastText to handle out-of-vocabulary words – an important advantage in the processing of German compound and inflected forms.

FastText's focus on computational efficiency was another factor why we chose this model. The authors state that the models' performance is on par with LSTM- and RNN-based models while being an order of magnitude faster to compute. For an application in robotics, this should be considered since our mobile robots have a limited amount of computation and energy resources. Therefore we chose FastText as a baseline model. However, an important shortcoming of FastText is, that it is not context-aware. With this model, every word is represented with the same vector, independent of the context of the word.

To train the FastText models, we first trained a skip-gram word vector with the length 100 on the texts of all 5.4 million samples. Both models have been trained 20 epochs,

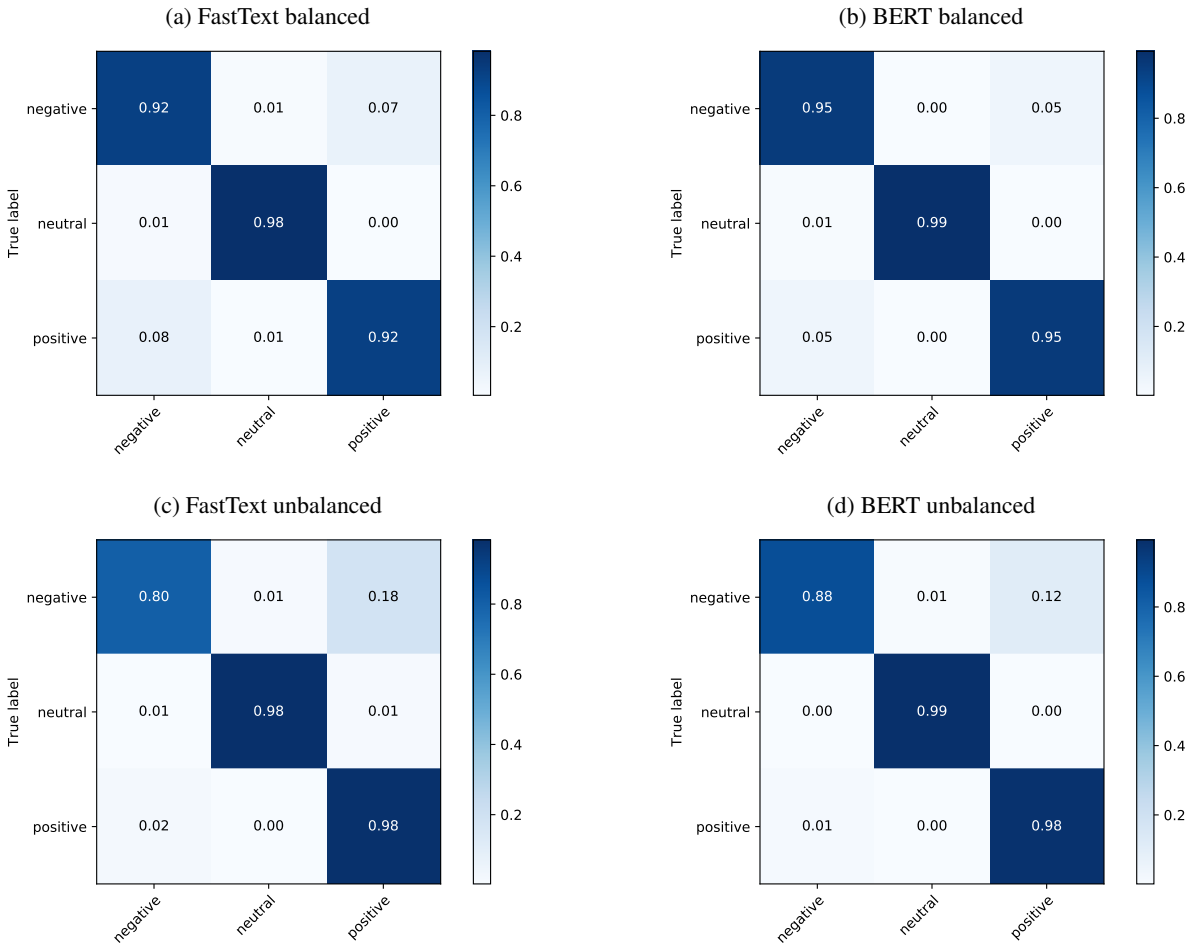

Figure 1: Four confusion matrices to compare FastText and BERTs classification results on the balanced and unbalanced data set. For both models the accuracy on the negative class drops when trained with the unbalanced sets. Note that all values are rounded.

with a learning rate of 0.1, word n-grams set to two and our pre-trained word vector.

Data Set	Balanced	Unbalanced
Scare	0.9071	**0.9083**
GermEval-2017	0.6970	**0.6980**
holidaycheck	0.9296	**0.9639**
SB10k	**0.6862**	0.6213
Filmstarts	0.8206	**0.8432**
PotTS	0.5268	**0.5416**
Emotions	**0.9913**	0.9773
leipzig-wikipedia	0.9883	**0.9886**
combined	0.9405	**0.9573**

Table 2: Micro averaged F_1 scores for FastText trained on the balanced and unbalanced data set.

Table 2 shows the micro F_1 scores for the balanced and unbalanced data set. The combined micro F_1 score is with 0.9573 to 0.9405 higher on the unbalanced data set.
We also calculated the macro F_1 score, with 0.9406 for the balanced and 0.9268 for the unbalanced data set. The difference between the micro and macro F_1 score indicates that the model, trained on the unbalanced data set, is biased towards a class with more samples. Comparing the confusion matrices for the models of both data sets in figure 1a and 1c shows, that the accuracy of the negative class drops from 0.92 to 0.80. These 12% of the negative samples were mostly labelled as positive. Therefore, we prefer the FastText model trained on the balanced data set.

4.2. BERT

A more recent model that is able to create context-depended word representations is *BERT*. This model uses bidirectional training of a deep transformer-based network architecture (Vaswani et al., 2017). From this architecture, however, BERT uses only the transformer-encoder layer, which consists of a scaled dot-product attention layer and a feed-forward neural network layer. The authors of BERT have released pre-trained models of two different sizes, namely *BERT small* which consists of 12 transformer-encoder layers, and *BERT large* which consists of 24 of these layers.
For our work, we have decided to use the BERT implementation[4] described in (Wolf et al., 2019). The repository also provides a German BERT small model.
For the sentiment classification task, the output of BERT

[4]https://github.com/huggingface/transformers

is fed into a feed-forward neural network. The weights of BERT's transformer layers are initialized with the pre-trained model and jointly trained with the feed-forward classification layer.

Following the recommendations from (Devlin et al., 2019), we trained all BERT models with a learning rate of $2 \cdot 10^{-5}$, a batch size of 32 and a maximal token length of 256 for three epochs. We did not perform an extensive hyperparameter optimization. Doing so might improve the models' results.

Table 3 shows that the BERT model trained on the unbalanced data set achieves +1.08 F_1 score compared to the balanced data set. As the FastText model, the BERT model trained on the unbalanced data set is biased towards the positive class. Figures 1b and 1d show, that the accuracy of the negative class drops from 0.95 to 0.88 and that the number of negative samples classified as positive raises from 5% to 12%. Therefore we recommend training BERT models on the balanced data set. Overall BERT outperforms FastText by +2.31 F_1 on the balanced data set.

However, the computational costs of BERT are higher than FastTexts. To classify 124194 samples BERT takes 11 minutes running on a Nividia 2080 Ti GPU, FastText classifies the identical data set in 5 seconds on an 8 core Intel CPU.

Data Set	Balanced	Unbalanced
Scare	0.9409	**0.9436**
GermEval-2017	0.7727	**0.7885**
holidaycheck	0.9552	**0.9775**
SB10k	**0.6930**	0.6720
Filmstarts	0.9062	**0.9219**
PotTS	0.6423	**0.6502**
Emotions	**0.9652**	0.9621
leipzig-wikipedia	**0.9983**	0.9981
combined	0.9636	**0.9744**

Table 3: Micro averaged F_1 scores for BERT trained on the balanced and unbalanced data set.

5. Ablation Studies

An important question with machine learning models is: Can the trained model generalize and successfully apply the learned concepts to an unseen domain? To approach this question, we left one data set out from our combined data set and retrained all models. We evaluated the resulting models with the test set and the test set of the data set we left out. We choose the *Scare* data set for this study. (Sänger et al., 2016) state that the data set consists of short app reviews written in a colloquial language that is more similar to twitter conversations than other reviews. By leaving out this data set, we can evaluate if the models were able to learn information from both domains and apply them to this data set. Table 4 compares the results from BERT and FastText trained on this task. Although the F_1 scores of both models dropped on the unseen *Scare* set, both models learned features from the combined data set that they could apply to the unseen *Scare* app reviews. However, BERT outperformed FastText by 9.4% achieving a F_1 score of 0.80 compared to FastText's F_1 score of 0.73.

Data Set	FastText	BERT	Gain%
GermEval-2017	0.6843	0.7727	+12.9
holidaycheck	0.9275	0.9542	+2.9
SB10k	0.6626	0.6840	+3.2
Filmstarts	0.8414	0.9164	+8.9
PotTS	0.5816	0.6667	+14.6
Emotions	0.9916	0.9832	-0.8
leipzig-wikipedia	0.9867	0.9985	+1.2
combined	0.9415	0.9649	+2.5
Scare	**0.7338**	**0.8039**	**+9.4**

Table 4: Micro averaged F_1 scores for FastText and BERT trained on the balanced data set without *Scare*. The last row contains the models scores on the *Scare* test data. The last column contains the gain / loss that BERT provides over FastText.

6. Conclusion

In this work, we want to present an approach to a broad-coverage sentiment classification model for the German language. We have combined different manually labelled data with large user labelled data sets. The resulting data set contains different domains, covering various language styles. With this approach we were able to create the largest publicly available German corpus for sentiment classification, containing more than 5,3 million examples. Furthermore, compared the performance of two different classification models on our data set. Both models performed well, with a F_1 score of 0.9405 for FastText and 0.9636 for BERT. We trained both models on a balanced and unbalanced version of our data set. The resulting models achieved both lower accuracy on the negative class. Therefore we recommend training these models on the balanced data set. We also tested how both models perform on data from unseen domains. In this test, BERT scored about 9% better than FastText. This indicates, that BERT is more suitable to classify data from unseen domains.

In our opinion, it is worth to further diversify our data set by adding more data from different domains. For this reason, we made our data and the associated source code available.

7. Acknowledgement

This research has been funded by the European Social Fund (ESF), SAB grant number 100339497.

8. Bibliographical References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*, June. arXiv: 1607.04606.

Chen, C.-Y., Yu, D., Wen, W., Yang, Y. M., Zhang, J., Zhou, M., Jesse, K., Chau, A., Bhowmick, A., Iyer, S., and others. (2018). Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data. *2nd Proceedings of Alexa Prize*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Fang, H., Cheng, H., Sap, M., Clark, E., Holtzman, A., Choi, Y., Smith, N. A., and Ostendorf, M. (2018). Sounding Board: A User-Centric and Content-Driven Social Chatbot. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, New Orleans, Louisiana. Association for Computational Linguistics.

Hellbach, S., Bahrmann, F., Donner, M., Himstedt, M., Klingner, M., Fonfara, J., Poschmann, P., Schmidt, R., and Böhme, H.-J. (2013). Learning as an essential ingredient for a tour guide robot. In *Workshop New Challenges in Neural Computation 2013*, page 53. Citeseer.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *arXiv:1607.01759 [cs]*, August. arXiv: 1607.01759.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

Naderalvojoud, B., QasemiZadeh, B., and Kallmeyer, L. (2017). HU-HHU at GermEval-2017 Sub-task B: Lexicon-Based Deep Learning for Contextual Sentiment Analysis. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Meia Customer Feedback*, pages 18–21, Berlin, Germany.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, July.

Poschmann, P., Donner, M., Bahrmann, F., Rudolph, M., Fonfara, J., Hellbach, S., and Böhme, H.-J. (2012). Wizard of Oz revisited: Researching on a tour guide robot while being faced with the public. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 701–706. IEEE.

Schmitt, M., Steinheber, S., Schreiber, K., and Roth, B. (2018). Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114, Brussels, Belgium. Association for Computational Linguistics.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*, December. arXiv: 1706.03762.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*, October. arXiv: 1910.03771.

Zhang, X., Zhao, J., and LeCun, Y. (2016). Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*, April. arXiv: 1509.01626.

9. Language Resource References

Cieliebak, Mark and Deriu, Jan Milan and Egger, Dominic and Uzdilli, Fatih. (2017). *A Twitter Corpus and Benchmark Resources for German Sentiment Analysis*. Association for Computational Linguistics.

Goldhahn, Dirk and Eckart, Thomas and Quasthoff, Uwe. (2012). *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages*. European Language Resources Association (ELRA).

Remus, R., Quasthoff, U., and Heyer, G. (2010). SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Language Ressources and Evaluation (LREC'10)*, pages 1168–1171.

Sidarenka, Uladzimir. (2016). *PotTS: The Potsdam Twitter Sentiment Corpus*. European Language Resources Association (ELRA).

Sänger, Mario and Leser, Ulf and Kemmerer, Steffen and Adolphs, Peter and Klinger, Roman. (2016). *SCARE — The Sentiment Corpus of App Reviews with Fine-grained Annotations in German*. European Language Resources Association (ELRA).

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., and Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, 41(2):534–538, May.

Waltinger, U. (2010). GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May. electronic proceedings.

Wojatzki, Michael and Ruppert, Eugen and Holschneider, Sarah and Zesch, Torsten and Biemann, Chris. (2017). *GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*.