A Twitter Corpus and Benchmark Resources for German Sentiment Analysis

Conference Paper · April 2017

DOI: 10.18653/v1/W17-1106

4 authors, including:

Mark Cieliebak
Zurich University of Applied Sciences
45 PUBLICATIONS **608** CITATIONS
SEE PROFILE

Jan Deriu
Zurich University of Applied Sciences
10 PUBLICATIONS **36** CITATIONS
SEE PROFILE

A Twitter Corpus and Benchmark Resources for German Sentiment Analysis

Mark Cieliebak
SpinningBytes
mc@spinningbytes.com

Jan Deriu
Zurich University of Applied Sciences
deri@zhaw.ch

Dominic Egger
Zurich University of Applied Sciences
eggo@zhaw.ch

Fatih Uzdilli
Zurich University of Applied Sciences
uzdi@zhaw.ch

Abstract

In this paper we present SB10k, a new corpus for sentiment analysis with approx. 10,000 German tweets.
We use this new corpus and two existing corpora to provide state-of-the-art benchmarks for sentiment analysis in German: we implemented a CNN (based on the winning system of SemEval-2016) and a feature-based SVM and compare their performance on all three corpora.
For the CNN, we also created German word embeddings trained on 300M tweets. These word embeddings were then optimized for sentiment analysis using distant-supervised learning.
The new corpus, the German word embeddings (plain and optimized), and source code to re-run the benchmarks are publicly available.

1 Introduction

With the advance of deep learning in text analytics, many benchmarks for text analytics tasks have been significantly improved in the last four years. For this reason, Zurich University of Applied Sciences (ZHAW) and SpinningBytes AG are collaborating in a joint research project to develop state-of-the-art solutions for text analytics tasks in several European languages. The goal is to adapt and optimize algorithms for tasks like sentiment analysis, named entity recognition (NER), topic extraction etc. into industry-ready software libraries.

One very challenging task is automatic sentiment analysis. The goal of sentiment analysis is to classify a text into the classes positive, negative, mixed, or neutral. Interest in automatic sentiment analysis has recently increased in both academia and industry due to the huge number of documents which are publicly available on social media. In fact, there exist various initiatives in the scientific community (such as shared tasks at SemEval (Nakov et al., 2016) or TREC (Ounis et al., 2008)), competitions at Kaggle[1], special tracks at major conferences like EMNLP or LREC, and several companies have built commercial sentiment analysis tools (Cieliebak et al., 2013).

Deep learning for sentiment analysis. Deep neural networks have become very successful for sentiment analysis. In fact, the winner and many top-ranked systems in SemEval-2016 were using deep neural networks (SemEval is an international competition that runs every year several tasks for semantic evaluation, including sentiment analysis) (Nakov et al., 2016). The winning system uses a multi-layer convolutional neural network that is trained in three phases. For English, this system achieves an F1-score of 62.7% on the test data of SemEval-2016 (Deriu et al., 2016), and top scores on test data from previous years. For this reason, we decided to adapt the system for sentiment analysis in German. Details are described in Section 4.

A new corpus for German sentiment. In order to train the CNN, millions of unlabeled and weakly-labeled German tweets are used for creating the word embeddings. In addition, a sufficient amount of manually labeled tweets is required to train and optimize the system. For languages such as English, Chinese or Arabic, there exist plenty of labeled training data for sentiment analysis, while for other European languages, the resources are often very limited (cf. "Related Work"). For German, in particular, we are only aware of three sentiment corpora of significant size: the DAI tweet data set, which contains 1800 German tweets with tweet-level sentiments (Narr et al., 2012); the

[1] https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

MGS corpus, which contains 109,130 German tweets (Mozetič et al., 2016); and the PotTS corpus, which contains 7992 German tweets that were annotated on phrase level (Sidarenka, 2016). Unfortunately, the first corpus is too small for training a sentiment system, the the second corpus has a very low inter-annotator agreement ($\alpha = 0.34$), indicating low-quality annotations, and the third corpus is not on sentence level.

For this reason, we decided to construct a large sentiment corpus with German tweets, called *SB10k*. This corpus should allow to train high-quality machine learning classifiers. It contains 9783 German tweets, each labeled by three annotators. Details of corpus construction and properties are described in Section 3.

Benchmark for German Sentiment. We evaluate the performance of the CNN on the three German sentiment corpora CAI, MGS, and SB10k in Section 5. In addition, we compare the results to a baseline system, a feature-based Support Vector Machine (SVM). To our knowledge, this is the first large-scale benchmark for sentiment analysis on German tweets.

Main Contributions. Our main contributions are:

- Benchmarks for sentiment analysis in German on three corpora.
- A new corpus *SB10k* for German sentiment with approx. 10000 tweets, manually labeled by three annotators.
- Publicly available word embeddings trained on 300M million German tweets (using word2vec), and modified word embeddings after distant-supervised learning with 40M million weakly-labeled sentiment tweets.

The new corpus, word embeddings for German (plain and fully-trained) and source code to re-run the benchmarks are available at `www.spinningbytes.com/resources`.

2 Related Work

There exists a tremendous amount of literature on sentiment analysis in general; for a good introduction and overview, see the recent book by Bing Liu (Zhao et al., 2016).

Corpora. Several human labeled corpora for sentiment analysis are available, which differ in: languages they cover, size, annotation schemes (number of annotators, sentiment), and document domains (tweets, news, blogs, product reviews etc.). For English there exist various corpora, e.g. for tweets (Narr et al., 2012), product reviews (Hu and Liu, 2004) or news (Wiebe et al., 2005), and sentiment corpora exist also for other European languges such as Italian (Straniscim et al., 2016), French (Bosco et al., 2016), Spanish (Martinez-Camara et al., 2016; Martinez-Camara et al., 2015) or Dutch (Verhoeven and Daelemans, 2014).

Sentiment Analysis in German. German is the most-spoken native language in Europe[2], and several research activities and events are focussed on German sentiment analysis. The Interest Group on German Sentiment Analysis (IGGSA)is a European collaboration of researchers working on German sentiment analysis. Among other things, they hosted several workshops and shared tasks on German Sentiment analysis, e.g. GESTALT-2014 (Ruppenhofer et al., 2014). For an extended list of publications on sentiment analysis in German, we refer the reader to IGGSA[3] .

Machine Learning for Sentiment Analysis. Until recently, feature-based systems were frequently used for sentiment analysis. In fact, almost all systems participating in SemEval-2014 were feature-based, with SVM, MaxEnt, and Naive Bayes being the most popular classifiers in the competition (Rosenthal et al., 2014). However, neural networks have shown great promise in NLP over the past few years. Examples are in semantic analysis (Shen et al., 2014), machine translation (Gao et al., 2014) and sentiment analysis (Socher et al., 2013). In particular, shallow convolutional neural networks (CNNs) have recently improved the state-of-the-art in text polarity classification demonstrating a significant increase in terms of accuracy compared to previous state-of-the-art techniques (Kim, 2014; Kalchbrenner et al., 2014; dos Santos and Gatti, 2014; Severyn and Moschitti, 2015; Johnson and Zhang, 2015; Rothe et al., 2016; Deriu et al., 2017).

[2] www.languageknowledge.eu

[3] https://sites.google.com/site/iggsahome/

3 Corpus Construction

3.1 Goals

We constructed a new sentiment corpus with German tweets, called *SB10k*. This corpus should allow to train high-quality machine learning classifiers. Based on our experiences with machine learning in other languages, we aimed at the following goals:

- The corpus should contain 10000 tweets, to provide sufficient data for complex system to be trained
- Selected tweets should cover a wide variety of unigrams and topics
- Each tweet should be labeled by three expert annotators
- Sentiment labels should be as balanced as possible

3.2 Basic Data Set

Our initial data was made up of tweets collected between 01.08.2013 and 31.10.2013. Those tweets were a random subselection (10%) of all tweets published during that time span. With the langid.py tool (Lui and Baldwin, 2012) we selected all German tweets from within our initial data. To minimize false positives, we only included tweets with a German confidence score of over 0.999. This resulted in 5.280.157 tweets.

3.3 Tweet Selection

Next we selected the tweets to be annotated. In order to achieve a large variety of topic and unigrams that are covered by the corpus, we applied a k-means clustering with bag of words features and cosine similarity to create 2500 clusters of tweets. Our goal was to have - at the end - four tweets per cluster, one for each sentiment class.

The majority of tweets in Twitter do not contain any opinion at all. Hence, selecting a random set of tweets for manual annotation would result in an unbalanced set, with a strong majority of neutral tweets. To find tweets with potentially different sentiments, we used a straight-forward approach: For each tweet we counted the number of positive and negative polarity words in per tweet, using the German polarity clues lexicon (Waltinger, 2010). Using these polarity words as indicators, we selected tweets that were "probably" positive, negative, mixed, or neutral: A tweet was considered "probably positive" if it contained at least one positive polarity words, but no negative polarity words; "probably negative" analogously; "probably mixed" if both types of polarity words occured; and "probably neutral" if no polarity words occured. In order to reach an as balanced corpus as possible and increase the number of tweets with an opinion, we decided to use primarily probably mixed tweets, since they tended to be anything but neutral. Obviously, this approach lessened the number of observed unigrams and topics to some degree.

3.4 Manual Annotation

We had 34 annotators (students in computer science or linguistics). Every tweet was shown to 3 random annotators and labeled with a sentiment class by each of those. They were given several examples and instructed to "categorize the sentiment expressed in a tweet, not the sentiment felt when reading the tweet". We added a non-German flag to clean out tweets wich slipped by the language identification, and tweets were marked as "unknown" when annotators could not decide on its sentiment.

3.5 Corpus Properties

Basic Outline. The corpus *SB10k* contains 9783 German tweets. Each tweet has sentiment annotations on tweet level by 3 human annotators, using sentiment classes positive, negative, neutral, mixed, and unknown. We aggregate the annotators' individual classes to assign a sentiment to each tweet, where tweet t has sentiment S if at least 2 annotators marked the tweet with S; otherwise, sentiment of t is unknown. The distribution of aggregated annotations is shown in Table 1.

Pos.	Neg.	Neutral	Mixed	Unknown	**Total**
1682	1077	5266	330	1428	9738

Table 1: Number of tweets per sentiment in *SB10k*

Unigram Diversity. Goal of our clustering approach was to achieve a high diversity of unigrams in our corpus. We therefore compare the diversity of the tweets that were selected by our clustering versus randomly sampled tweets. There are $u = 11.592.947$ distinct unigrams in all collected German tweets (approx. 5 million). There are 9452 unigrams in the labeled tweets (picked from

the k-means clustering), thus, the corpus covers 0.00081% of all unigrams. To compare this value to random sampling, we randomly picked 10000 tweets from all available tweets. This was repeated 10 times, resulting in an average coverage of 0.00075% of all unigrams. Thus, our clustering approach increases the number of encountered unigrams by 10.7%.

Annotator Agreement. To analyze the inter-annotator agreement within our corpus, we use Krippendorffs Alpha-reliability (Krippendorff, 2007). This agreement score fits well with our annotation scheme, in contrast to other scores like Kohens Kappa, since Krippendorffs Alpha basically computes the coincidence matrix between any two annotators, and calculates a weighed sum. We had pairs of annotators which shared as little as 1 tweet and pairs which shared as many as 1673 tweets. To mitigate this issue, we only considered pairs of annotators which shared at least 50 tweets. This results in $\alpha = 0.39$, with a standard deviation of 0.12.

4 Benchmark System: Multi-layer CNN with Three-Phase Training

4.1 Architecture and Implementation

The winning system of SemEval-2016 by team "SwissCheese" is based on a convolutional neural network (CNN) which is trained in three phases. We adapted and optimized the system for German sentiment analysis. In the following, we briefly describe the high-level architecture and parameters of this CNN. For more details on the network topology and technical architecture, see citederiu17www.

The core component of the system is a multi-layer convolutional neural network (CNN), which consists in two consecutive pairs of convolutional-pooling layers, followed by a single fully connected hidden layer and a soft-max output layer. The system is trained in three phases. Figure 1 shows a complete overview of the phases of the learning procedure: i) unsupervised phase, where word embeddings are created on a corpus of 300M unlabeled tweets; ii) distant supervised phase, where the network is trained on a weakly-labeled dataset of 40M tweets containing emoticons; and iii) supervised phase, where the network is nally trained on manually annotated tweets. For English, a similar system achieved an F1-score of

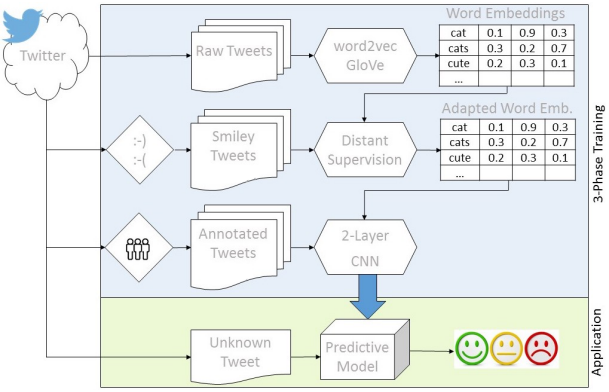

Figure 1: *Training Phases Overview.*

62.7% on the test data of SemEval-2016 (Deriu et al., 2016).

Training. The word embeddings are learned on an unsupervised corpus containing 300M German tweets. We apply a skip-gram model of window-size 5 and filter words that occur less than 15 times (Severyn and Moschitti, 2015). The dimensionality of the vector representation is set to $d = 52$. During the distant-supervised phase, we use emoticons to infer noisy labels on the tweets in the training set (Read, 2005; Go et al., 2009). We used 40M tweets (8M negative, 32M positive). The neural network was trained on these data for one epoch, before finally training on the supervised data for about 20 epochs. The word-embeddings are updated during both the distant- and the supervised training phases by applying back-propagation through the entire network.

Computing Time for Training. On a GPU computer with 3072 cores and 8GB of RAM, it took approximately 24 hours to create the word embeddings, 15 hours for the distant-supervised phase, and 30 minutes for the supervised phase.

5 Benchmark for German Sentiment Analysis

We now study how the CNN performs when trained and/or tested on the three German sentiment corpora we are aware of: *SB10k* (from this paper, 9738 tweets), *MGS* corpus (109'130 tweets, (Mozetič et al., 2016)), and *DAI* corpus (1800 tweets, (Narr et al., 2012)). Corpora *SB10k* and *MGS* were randomly split into training (90%) and

Classifier	Training Corpus	Test Corpus	$F1_{pos}$	$F1_{neg}$	$F1_{neutral}$	F1
SVM	SB10k	SB10k	66.16	47.80	81.32	56.98
CNN	SB10k	SB10k	71.46	58.72	81.19	**65.09**
SVM	SB10k	MGS	49.50	38.62	66.41	44.06
CNN	SB10k	MGS	50.41	44.19	71.81	**47.30**
SVM	SB10k	DAI (full)	62.30	61.40	81.22	**61.85**
CNN	SB10k	DAI (full)	62.79	58.43	79.92	60.61
SVM	MGS	SB10k	67.77	53.23	80.20	60.50
CNN	MGS	SB10k	63.94	58.21	70.66	**61.07**
SVM	MGS	MGS	60.34	56.48	69.31	58.41
CNN	MGS	MGS	61.49	58.12	68.62	**59.80**
SVM	MGS	DAI (full)	59.32	56.03	74.83	57.68
CNN	MGS	DAI (full)	61.01	55.74	76.88	**58.38**

Table 2: Benchmarks for sentiment in German. SVM and CNN were trained on fixed split of each corpus (90%), and then tested on the remaining texts. For DAI, all texts were used for testing. F1 is macroaveraged from $F1_{pos}$ and $F1_{neg}$. Bold numbers identify higher F1 score of both classifiers for each combination of test and training corpus (2 lines).

testing (10%) subsets[4]. DAI was not split, since it was only used for testing.

For comparison, we implemented a feature-based system using a Support Vector Machine (SVM). Feature selection is based on the system described in (Uzdilli et al., 2015), which ranked 8th in the Semeval competition of 2015, and include n-gram, various lexical features, and statistical text properties. We use the macro-averages F1-score of positive and negative class, i.e. F1 = ($F1_{pos}$ + $F1_{neg}$) / 2, since this is also used in SemEval (Rosenthal et al., 2015) as a standard measure of quality. The results are reported in Table 2.

Results. We observe from Table 2 that CNN outperforms SVM in all but one case (SB10k-DAI). Surprisingly, SVM performs better on SB10k when trained on the foreign corpus MGS then when trained on SB10k (60.50 instead of 56.98), while in all other cases the classifier benefits when being trained on the same corpus. There is a high variance in F1-score for the same system on different test corpora, e.g. between 47.30 and 65.09 for CNN trained on SB10k.

Both SVM and CNN outperform the reference system from (Mozetič et al., 2016), which reported an F1-score of 53.6 for the German part of *MGS* (note that they used cross-validation instead of a fixed split of the corpus).

[4]These splits are available at `www.spinningbytes.com/resources` to allow other researchers to compare their results with the benchmarks

We also computed macroaveraged 3-class F1-score $F1_3$ = ($F1_{pos}$ + $F1_{neg}$ + $F1_{neutral}$) / 3, which is on average 4.42 points higher than F1, due to the higher values of $F1_{neutral}$.

6 Conclusion

We have evaluated two state-of-the-art systems for sentiment analysis in German on three Twitter corpora (on of them new). Since all corpora are publicly available, these results can serve as a benchmark for other sentiment systems in German.

The results show that the deep learning system outperforms the feature-based system in all but one cases. However, F1-score is around 60% in most cases, even when a system is trained and tested on the same corpus (with a fixed split of data). This means that there is still potential for inprovement.

7 Acknowledgements

This research has been funded by Commission for Technology and Innovation (CTI) project no. 18832.1 PFES-ES and by SpinningBytes AG, Switzerland. We would like to thank Leon Derczynski for giving us access to his tweet collection, and Matthias Flour for his support in labeling the tweets.

References

Cristina Bosco, Mirko Lai, Viviana Patti, and Daniela Virone. 2016. Tweeting and Being Ironic in the Debate about a Political Reform: the French Annotated Corpus TWitter-MariagePourTous. In *Proceedings of LREC-2016*, pages 1619–1626, Portoroz, Slovenia.

Mark Cieliebak, Oliver Dürr, and Fatih Uzdilli. 2013. Potential and Limitations of Commercial Sentiment Detection Tools. In *Proceedings of ESSEM@AI*IA*, pages 47–58, Torino, Italy.

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. SwissCheese at SemEval-2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of SemEval-2016*, pages 1124–1128, San Diego, California, USA.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging large amounts of weakly supervised data for multilanguage sentiment classification. In *Proceedings of WWW-2017*, Peth, Australia.

Cícero Nogueira dos Santos and Maira Gatti. 2014. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Proceedings of COLING-2014*, pages 69–78, Dublin, Ireland.

Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2014. Learning continuous phrase representations for translation modeling. In *Proceedings of ACL-2014*, pages 699–709, Baltimore, Maryland, USA.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision. Technical report, The Stanford Natural Language Processing Group, Stanford, CA, USA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD*, pages 168–177, Seattle, Washington, USA.

Rie Johnson and Tong Zhang. 2015. Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. In *Advances in Neural Information Processing Systems 28*, pages 919–927, Montreal, Canada.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of ACL-2014*, pages 655–665, Baltimore, Maryland, USA.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of EMNLP-2014*, pages 1746–1751, Doha, Quatar.

Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)*, page 43.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30, Jeju Island, Korea.

Eugenio Martinez-Camara, M. Teresa Martin-Valdivia, L. Alfonso Urena-Lopez, and Ruslan Mitkov. 2015. Polarity classification for Spanish tweets using the COST corpus. *Journal of Information Science*, 41(3):263–272.

Eugenio Martinez-Camara, Miguel A. Garcia-Cumbreras, Julio Villena-Roman, and Janine Garcia-Morera. 2016. TASS 2015 - The Evolution of the Spanish Opinion Mining Systems. *Procesamiento del Lenguaje Natural*, 56:33–40.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PloS one*, 11(5):e0155036.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of SemEval-2016*, pages 1–18, San Diego, USA.

Sascha Narr, Michael Hulfenhaus, and Sahin Albayrak. 2012. Language-independent twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML), LWA*, pages 12–14.

Iadh Ounis, Craig Macdonald, and Ian Soboroff. 2008. Overview of the TREC-2008 Blog Track. In *Proceedings of TREC-2008*.

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pages 43–48, Ann Arbor, Michigan.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of SemEval-2014*, pages 73 – 80, Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. *Proceedings of SemEval-2015*, pages 451–463.

Sascha Rothe, Sebastian Ebert, and Hinrich Schutze. 2016. Ultradense Word Embeddings by Orthogonal Transformation. *arXiv*.

Josef Ruppenhofer, Roman Klinger, Julia Maria Struß, Jonathan Sonntag, and Michael Wiegand. 2014. IGGSA Shared Tasks on German Sentiment Analysis (GESTALT). In *Workshop Proceedings of the 12th*

Edition of the KONVENS Conference, pages 164 – 173, Hildesheim, Germany.

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of SemEval-2015*, pages 464–469, Denver, Colorado.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 101–110, Shanghai, China.

Uladzimir Sidarenka. 2016. PotTS: The Potsdam Twitter Sentiment Corpus. In *Proceedings of LREC-2016*, pages 1133 – 1141, Portoroz, Slovenia.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP-2013*, volume 1631, page 1642, Seattle, Washington, USA.

Marco Straniscim, Christina Bosco, Delia Irazu Hernnandez Farias, and Viviana Patti. 2016. Annotating Sentiment and Irony in the Online Italian Political Debate on labuonascuola. In *Proceedings of LREC-2016*, pages 2892–2899, Portoroz, Slovenia.

Fatih Uzdilli, Martin Jaggi, Dominic Egger, Pascal Julmy, Leon Derczynski, and Mark Cieliebak. 2015. Swiss-Chocolate: Combining Flipout Regularization and Random Forests with Artificially Built Subsystems to Boost Text-Classification for Sentiment. In *Proceedings of SemEval-2015*, pages 608–612, Denver, Colorado.

Ben Verhoeven and Walter Daelemans. 2014. CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of LREC-2014*, pages 3081–3085, Reykjavik, Iceland.

Ulli Waltinger. 2010. GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In *Proceedings of LREC-2010*, pages 1638–1642, Valletta, Malta.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Jun Zhao, Kang Liu, and Liheng Xu. 2016. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. MIT Press, Cambridge, MA, USA.