Webis: An Ensemble for Twitter Sentiment Detection

Matthias Hagen **Martin Potthast** **Michel Büchner** **Benno Stein**
Bauhaus-Universität Weimar
<first name>.<last name>@uni-weimar.de

Abstract

We reproduce four Twitter sentiment classification approaches that participated in previous SemEval editions with diverse feature sets. The reproduced approaches are combined in an ensemble, averaging the individual classifiers' confidence scores for the three classes (positive, neutral, negative) and deciding sentiment polarity based on these averages. The experimental evaluation on SemEval data shows our re-implementations to slightly outperform their respective originals. Moreover, not too surprisingly, the ensemble of the reproduced approaches serves as a strong baseline in the current edition where it is top-ranked on the 2015 test set.

1 Introduction

We reproduce four state-of-the-art approaches to classifying the sentiment expressed in a given tweet, and combine the four approaches to an ensemble based on the individual classifiers' confidence scores. In particular, we focus on subtask B of SemEval 2015's task 10 "Sentiment Analysis in Twitter," where the goal is to classify the whole tweet as either positive, neutral, or negative. Since the notebook descriptions accompanying submissions to shared tasks are understandably very terse, it is often a challenge to reproduce the results reported. Therefore, we attempt to reproduce the state-of-the-art Twitter sentiment detection algorithms that have been submitted to the aforementioned task in its previous two editions. Furthermore, we combine the reproduced classifiers in an ensemble. Since the individual approaches employ diverse feature sets, the goal of the ensemble is to combine their individual strengths. The paper at hand is a slight extension of the approach from our ECIR 2015 reproducibility track paper (Hagen et al., 2015) such that also text passages are reused. In our ECIR paper, we showed that three selected approaches participating in the SemEval 2013 Twitter sentiment task 2 could be reproduced from the papers accompanying the individual approaches. Adding the best participant of the respective SemEval 2014 task 9 is shown to form a very strong baseline that was not outperformed by the SemEval 2015 participants on the 2015 test data and that also places in the top-10 in the progress test.

In Section 2 we briefly describe some related work while in Section 3 we provide more details on the four individual approaches as well as our ensemble scheme. Some concluding remarks and an outlook on future work close the paper in Section 4. An experimental evaluation of our approach and an in-depth comparison to the other participants is not included in this paper since it can be found in the task overview (Rosenthal et al., 2015).

2 Related Work

Sentiment detection is a classic problem of text classification. Unlike other text classification tasks, the goal is not to identify topics, entities, or authors of a text but to rate the expressed sentiment as positive, negative, or neutral. Most approaches used for sentiment detection usually involve methods from machine learning, computational linguistics, and statistics. Typically, several approaches from these fields are combined for sentiment detection (Pang et al., 2002; Turney, 2002; Feldman, 2013).

Since Twitter is one of the richest sources of opinion, a lot of different approaches to sentiment de-

tection in tweets have been proposed. Different approaches use different feature sets ranging from standard word polarity expressions or unigram features also applied in general sentiment detection (Go et al., 2009; Kouloumpis et al., 2011), to the usage of emoticons and uppercases (Barbosa and Feng, 2010), word lengthening (Brody and Diakopoulos, 2011), phonetic features (Ermakov and Ermakova, 2013), multi-lingual machine translation (Balahur and Turchi, 2013), or word embeddings (Tang et al., 2014). The task usually is to detect the sentiment expressed in a tweet as a whole (also focus of this paper). But it can also be used to identify the sentiment in a tweet with respect to a given target concept expressed in a query (Jiang et al., 2011). The difference is that a generally negative tweet might not say anything about the target concept and must thus be considered neutral with respect to the target concept.

Both tasks, namely sentiment detection in a tweet, and sentiment detection with respect to a specific target concept, are part of the SemEval sentiment analysis tasks since 2013 (Nakov et al., 2013; Rosenthal et al., 2014). SemEval fosters research on sentiment detection for short texts in particular, and gathers the best-performing approaches in a friendly competition. The problem we are dealing with is formulated as subtask B: given a tweet, decide whether its message is positive, negative, or neutral.

State-of-the-art approaches have been submitted to the SemEval tasks. However, up to now, no one had trained a meta-classifier based on the submitted approaches to determine what can be achieved when combining them, whereas each participating team only trains their individual classifier using respective individual feature sets. Our idea is to combine four of the best-performing approaches from the last years with different feature sets, and to form an ensemble classifier that leverages the individual classifiers' strengths forming a strong baseline.

Ensemble learning is a classic approach of combining several classifiers to a more powerful ensemble (Opitz and Maclin, 1999; Polikar, 2006; Rokach, 2010). The classic approaches of Bagging (Breiman, 1996) and Boosting (Schapire, 1990; Freund and Schapire, 1996) try to either combine the outputs of different classifiers trained on different random instances of the training set or on training the classifiers on instances that were misclassified by the other classifiers. Both rather work on the final predictions of the classifiers just as for instance averaging or majority voting on the predictions (Asker and Maclin, 1997) would do. In our case, we employ the confidence scores of the participating classifiers. Several papers describe different ways of working with the classifiers' confidence scores, such as learning a dynamic confidence weighting scheme (Fung et al., 2006), or deriving a set cover with averaging confidences (Rokach et al., 2014). Instead, we simply average the three confidence scores of the three classifiers for each individual class. This straightforward approach performs superior to its individual parts and performs competitive in the SemEval competitions. Thus, its sentiment detection results can be directly used in any of the above use cases for Twitter sentiment detection.

3 Individual Approaches and Ensemble

For our ECIR 2015 reproducibility paper (Hagen et al., 2015), we originally selected three state-of-the-art approaches for Twitter sentiment detection among the 38 participants of SemEval 2013. To identify worthy candidates—and to satisfy the claim "state of the art"—we picked the top-ranked approach by team NRC-Canada (Mohammad et al., 2013). However, instead of simply picking the approaches on ranks two and three to complete our set, we first analyzed the notebooks of the top-ranked teams in order to identify approaches that are significantly dissimilar from NRC-Canada. We decided to handpick approaches this way so they complement each other in an ensemble. As a second candidate, we picked team GU-MLT-LT (Günther and Furrer, 2013) since it uses some other features and a different sentiment lexicon. As a third candidate, we picked team KLUE (Proisl et al., 2013), which was ranked fifth. We discarded the third-ranked approach as it is using a large set of not publicly available rules probably hindering reproducibility, whereas the fourth-ranked system seemed too similar to NRC and GU-MLT-LT to add something new to the planned ensemble. Finally, for participation in SemEval 2015, we also included TeamX (Miura et al., 2014) as the 2014 top-performing approach resulting in an ensemble of four.

Note that due to the selection process, reproducing the four approaches does not deteriorate into reimplementing the feature set of one approach and reusing it for the other two. Moreover, combining the four approaches into an ensemble classifier actually makes sense, since, due to the feature set diversity, they tap sufficiently different information sources. In what follows, we first briefly recap the features used by the individual classifiers and then explain our ensemble strategy.

3.1 NRC-Canada

Team NRC-Canada (Mohammad et al., 2013) used a classifier with a wide range of features. A tweet is first preprocessed by replacing URLs and user names by some placeholder. The tweets are then tokenized and POS-tagged. An SVM with linear kernel is trained using the following feature set.

N-grams The occurrence of word 1- to 4-grams as well as occurrences of pairs of non-consecutive words where the intermediate words are replaced by a placeholder. No term-weighting like $tf{\cdot}idf$ is used. Similarly for occurrence of character 3- to 5-grams.

ALLCAPS Number of all-capitalized words.

Parts of speech Occurrence of part-of-speech tags.

Polarity dictionaries In total, five polarity dictionaries are used. Three of these were manually created: the NRC Emotion Lexicon (Mohammad and Turney, 2010; Mohammad and Turney, 2013) with 14,000 words, the MPQA Lexicon (Wilson et al., 2005) with 8,000 words, and the Bing Liu Lexicon (Hu and Liu, 2004) with 6,800 words. Two other dictionaries were created automatically. For the first one, the idea is that several hash tags can express sentiment (e.g., #good). Team NRC crawled 775,000 tweets from April to December 2012 that contain at least one of 32 positive or 38 negative hash tags that were manually created (e.g., #good and #bad). For word 1-grams and word 2-grams in the tweets, PMI-scores were calculated for each of the 70 hash tags to yield a score for the n-grams (i.e., the ones with higher positive hash tag PMI are positive, the others negative). The resulting dictionary contains 54,129 unigrams, 316,531 bigrams, and 308,808 pairs of non-consecutive words. The second automatically created dictionary is not based on PMI for hash tags but for emoticons. It was created in a similar way as the hash tag dictionary and contains 62,468 unigrams, 677,698 bigrams, and 480,010 pairs of non-consecutive words.

For each entry of the five dictionaries, the dictionary score is either positive, negative, or zero. For a tweet and each individual dictionary, several features are computed: the number of dictionary entries with a positive score and the number of entries with a negative score, the sum of the positive scores and the sum of the negative scores of the tweet's dictionary entries, the maximum positive score and minimum negative score of the tweet's dictionary entries, and the last positive score and negative score.

Punctuation marks The number of non-single punctuation marks (e.g., !! or ?!) is used as a feature and whether the last one is an exclamation or a question mark.

Emoticons The emoticons contained in a tweet, their polarity, and whether the last token of a tweet is an emoticon are employed features.

Word lengthening The number of words that are lengthened by repeating a letter more than twice (e.g., cooooolll) is a feature.

Clustering Via unsupervised Brown clustering (Brown et al., 1992) a set of 56,345,753 tweets by Owoputi (Owoputi et al., 2013) clustered into 1,000 clusters. The IDs of the clusters in which the terms of a tweet occur are also used as features.

Negation The number of negated segments is a feature. A negated segment starts with a negation (e.g., shouldn't) and ends with a punctuation mark (Pang et al., 2002). Every token in a negated segment (words, emoticons) gets a suffix NEG attached (e.g., perfect_NEG).

3.2 GU-MLT-LT

Team GU-MLT-LT (Günther and Furrer, 2013) was ranked second in SemEval 2013. They train a stochastic gradient decent classifier on a much smaller feature set compared to NRC. The following feature set is computed for tokenized versions of the original raw tweet, a lowercased normalized version of the tweet, and a version of the lowercased tweet where consecutive identical letters are collapsed (e.g., helllo gets hello).

Normalized unigrams The occurrence of the normalized word unigrams is one feature set. No term weighting like for instance $tf \cdot idf$ is used.

Stems Porter stemming (Porter, 1980) is used to identify the occurrence of the stems of the collapsed word unigrams as another feature set. Again, no term weighting is applied.

Clustering Similar to NRC, the cluster IDs of the raw, normalized, and collapsed tokens are features.

Polarity dictionary The SentiWordNet assessments (Baccianella et al., 2010) of the individual collapsed tokens and the sum of all tokens' scores in a tweet are further features.

Negation Normalized tokens and stems are added as negated features similar to NRC.

3.3 KLUE

Team KLUE (Proisl et al., 2013) was ranked fifth in the SemEval 2013 ranking. Similarly to NRC, team KLUE first replaces URLs and user names by some placeholder and tokenizes the lowercased tweets. A maximum entropy-based classifier is trained on the following features.

N-grams Word unigrams and bigrams are used as features but in contrast to NRC and GU-MLT-LT not just by occurrence but frequency-weighted. Due to the short tweet length this however often boils down to a simple occurrence feature. To be part of the feature set, an n-gram has to be contained in at least five tweets. This excludes some rather obscure and rare terms or misspellings.

Length The number of tokens in a tweet (i.e., its length) is used as a feature. Interestingly, NRC and GU-MLT-LT do not explicitly use this feature.

Polarity dictionary The employed dictionary is the AFINN-111 lexicon (Nielsen, 2011) containing 2,447 words with assessments from -5 (very negative) to $+5$ (very positive). Team KLUE added another 343 words. Employed features are the number of positive tokens in a tweet, the number of negative tokens, the number of tokens with a dictionary score, and the arithmetic mean of the scores in a tweet.

Emoticons and abbreviations A list of 212 emoticons and 95 colloquial abbreviations from Wikipedia was manually scored as positive, negative, or neutral. For a tweet, again the number of positive and negative tokens from this list, the total number of scored tokens, and the arithmetic mean are used as features.

Negation Negation is not treated for the whole segment as NRC and GU-MLT-LT do but only on the next three tokens except the case that the punctuation comes earlier. Only negated word unigrams are used as an additional feature set. The polarity scores from the above dictionary are multiplied by -1 for terms up to 4 tokens after the negation.

3.4 TeamX

TeamX (Miura et al., 2014) was ranked first in the SemEval 2014 ranking. The approach was inspired by NRC Canada's 2013 method but uses fewer features and more polarity dictionaries—some differences are outlined below. Although it is very close to NRC Canada, some differences exist that justify TeamX's selection for our ensemble—besides its good performance in SemEval 2014.

Parts of speech Two different POS taggers are used: the Stanford POS tagger's tags are used for the polarity dictionaries based on formal language and for word sense disambiguation while the CMU ARK POS tagger is used for the polarity dictionaries containing more informal expressions, n-grams and the cluster features. Since the CMU ARK tagger was explicitly developed for handling tweets, it is better suited for the informal language often used in tweets while the Stanford tagger better addresses the needs of the formal dictionaries.

N-grams Word uni- up to 4-grams (consecutive words but also with gaps) and consecutive character 3- up to 5-grams are used as features similar to NRC.

Polarity dictionaries TeamX uses all the dictionaries of NRC, GU-ML-LT, and KLUE except for the NRC emoticon dictionary. Additionally, also SentiWordNet is used.

3.5 Remarks on Reimplementing

As was to be expected, it turned out to be impossible to re-implement all features precisely as the original authors did. Either not all data were publicly available, or the features themselves were not sufficiently explained in the notebooks. We deliberated to contact the original authors to give them a chance to supply missing data as well as to elaborate on

missing information. However, we ultimately opted against doing so for the following reason: our goal was to reproduce their results, not to repeat them. The difference between reproducibility and repeatability is subtle, yet important. If an approach can be re-implemented with incomplete information and if it then achieves a performance within the ballpark of the original, it can be considered much more robust than an approach that must be precisely the same as the original to achieve its expected performance. The former hints reproducibility, the latter only repeatability. This is why we have partly re-invented the approaches on our own, wherever information or data were missing. In doing so, we sometimes found ourselves in a situation where departing from the original approach would yield better performance. In such cases, we decided to maximize performance rather than sticking to the original, since in an evaluation setting, it is unfair to not maximize performance wherever one can.

In particular, the emoticons and abbreviations added by the KLUE team were not available, such that we only choose the AFINN-111 polarity dictionary and re-implemented an emoticon detection and manual polarity scoring ourselves. We also chose not to use the frequency information in the KLUE system but only Boolean occurrence like NRC and GU-MLT-LT, since pilot studies on the SemEval 2013 training and development sets showed that to perform much better. For all three approaches, we unified tweet normalization regarding lowercasing and completely removing URLs and user names instead of adding a placeholder. As for the classifier itself, we did not use the learning algorithms used originally but L2-regularized logistic regression from the LIBLINEAR SVM library for all three approaches. In our pilot experiments on the SemEval 2013 training and development set this showed a very good trade-off between training time and accuracy. We set the cost parameter to 0.5 for NRC, to 0.15 for GU-MLT-LT, and to 0.05 for TeamX and KLUE.

Note that most of our design decisions do not hurt the individual performances but instead improve the accuracy for GU-MLT-LT and KLUE on the SemEval 2013 test set. Table 1 shows the performance of the original SemEval 2013 and 2014 rankings and that of our re-implementations based on the averaged F1-score for the positive and negative class only (as is done at SemEval). While the reimplemented NRC performance is slightly better, GU-MLT-LT and KLUE are substantially improved. That TeamX lost performance is probably due to a fact that we only recognized after the competition: The word sense feature was unintentionally not switched on in the re-implementation of TeamX. Since for this "handicapped" version of TeamX (again, we just noticed the reason for the handicap after the SemEval 2015 deadline) the weighting scheme of the classification probabilities proposed for the original approach (Miura et al., 2014) did decrease the performance, we also did not use these weights. If we would have noticed our mistake before, the performance of the TeamX classifier would probably have been better.

Table 1: F1-scores of the original and reimplemented classifiers on the SemEval 2013 and 2014 test data and performance of the final system on the 2015 test data.

Classifier	**Original SemEval 2013**	**Reimplemented**
NRC	69.02	69.44
GU-MLT-LT	65.27	67.27
KLUE	63.06	67.05
	Original SemEval 2014	**Reimplemented**
TeamX	72.12	70.09
	SemEval 2015 result	
Ensemble	64.84	(rank 1 among 40 systems)

Altogether, we conclude that reproducing the SemEval approaches was generally possible but involved some subtleties that sometimes lead to difficult design decisions. Our resolution is to maximize performance rather than to dogmatically stick to the original approach; even though this includes the error in the TeamX re-implementation that went through unnoticed until after the deadline.

3.6 Ensemble Combination

In our pilot studies on the SemEval 2013 training and development sets, we tested several ways of combining the classifiers to an ensemble method. One of the main observations was that each individual approach classifies some tweets correctly that others fail for. This is not too surprising given the different feature sets but also supports

the idea of using an ensemble to combine the individual strengths. Although we briefly tried different ways of bagging and boosting the three classifiers, it soon turned out that some simpler combination performs better. A problem, for instance, was that some misclassified tweets are very difficult (e.g., the positive `Cant wait for the UCLA midnight madness tomorrow night`). Since often at least two classifiers fail on a hard tweet, this rules out some basic combination schemes, such as the majority vote which turned out to perform worse on the SemEval 2013 development set than NRC alone.

The solution that we finally came up with is motivated by observing how the classifiers trained on the SemEval 2013 training set behave for tweets in the development set. Typically, not the four final decisions but the respective confidences or probabilities of the individual classifiers give a good hint on uncertainties. If two are not really sure about the final classification, sometimes the remaining ones favor another class with high confidence. Thus, instead of looking at the classifications, we decided to use the confidence scores or probabilities to build the ensemble. This approach is also motivated by old and also more recent research on ensemble learning (Asker and Maclin, 1997; Fung et al., 2006; Rokach et al., 2014). But instead of learning a weighting scheme for the different individual classifiers, we decided to simply compute the average probability of the four classifiers for each of the three classes (positive, negative, neutral).

Our ensemble thus works as follows. The four individual re-implementations of the TeamX, the NRC, the GU-MLT-LT, and the KLUE classifier are individually trained on the SemEval 2013 training and development set as if being applied individually—without boosting or bagging. As for the classification of a tweet, the ensemble ignores the individual classifiers' classification decisions but requests the classifiers' probabilities (or confidences) for each class. The ensemble decision then chooses the class with the highest average probability—again, no sophisticated techniques like dynamic confidence weighting (Fung et al., 2006) or set covering schemes (Rokach et al., 2014) are involved. Thus, our final ensemble method is a rather straightforward system based on averaging confidences instead of voting schemes on the actual classifications of the individual classifiers. It can be easily implemented on top of the four classifiers and thus incurs no additional overhead. It also proves a very strong baseline in the SemEval 2015 evaluation. This is not really surprising since typically ensembles of good and diverse approaches should achieve better performances. Our code for the four reproduced approaches as well as that of the ensemble is publicly available.[1]

4 Conclusion and Outlook

We have reproduced four state-of-the-art approaches to sentiment detection for Twitter tweets. Our findings include that not all aspects of the approaches could be reproduced precisely, but that missing data, missing information, as well as opportunities to improve the approaches' performances lead us to re-invent them and to depart to some extent from the original descriptions. Most of our changes have improved the performances of the original approaches (except the erroneously and unintentionally switched off word sense feature of TeamX). Moreover, we have demonstrated that the approaches can be reproduced even with incomplete information about them, which is a much stronger property than being merely repeatable.

In addition, we investigated a combination of confidence scores of the four approaches within an ensemble that altogether yields a top-performing Twitter sentiment detection system forming a very strong baseline. The ensemble computation is as efficient as its components, and its effectiveness can be seen from the top rank on the SemEval 2015 test set and the top-10 ranking in the progress test involving the previous years' test data.

Promising directions for future research are an extensive error analysis and the identification of further classifiers potentially strengthening the ensemble. Following our philosophy of selecting approaches that are significantly different from each other, it will be interesting to observe how much new approaches can improve the existing ensemble.

[1] `http://www.uni-weimar.de/medien/webis/publications/by-year/#stein_2015d`

References

Lars Asker and Richard Maclin. 1997. Ensembles as a sequence of classifiers. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 23-29, 1997, 2 Volumes*, pages 860–865.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Alexandra Balahur and Marco Turchi. 2013. Improving sentiment analysis in twitter using multilingual machine translated data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2013*, pages 49–55.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 36–44.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 562–570.

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Sergei Ermakov and Liana Ermakova. 2013. Sentiment classification based on phonetic characteristics. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, pages 706–709.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Yoav Freund and Robert E. Schapire. 1996. Experiments with a new boosting algorithm. In *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996*, pages 148–156.

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Haixun Wang, David W. Cheung, and Huan Liu. 2006. A balanced ensemble approach to weighting classifiers for text classification. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China*, pages 869–873.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University.

Tobias Günther and Lenz Furrer. 2013. GU-MLT-LT: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 328–332.

Matthias Hagen, Martin Potthast, Michel Büchner, and Benno Stein. 2015. Twitter sentiment detection via ensemble classification using averaged confidence scores. In *Proceedings of ECIR 2015 (to appear)*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 151–160.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*.

Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. 2014. Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, CAAGET 2010*, pages 26–34.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, Greece, May 30, 2011*, pages 93–98.

David W. Opitz and Richard Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 380–390.

Bo Pang, Lillian Lee, and Shiuvakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2002*, pages 79–86.

Robi Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45.

Martin Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

Thomas Proisl, Paul Greiner, Stefan Evert, and Besim Kabashi. 2013. Klue: Simple and robust methods for polarity classification. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 395–401.

Lior Rokach, Alon Schclar, and Ehud Itach. 2014. Ensemble methods for multi-label classification. *Expert Systems with Applications*, 41(16):7507–7523.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval 2015*, SemEval 2015, Denver, Colorado, June. Association for Computational Linguistics.

Robert E. Schapire. 1990. The strength of weak learnability. *Machine Learning*, 5:197–227.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1555–1565.

Peter D. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002,J uly 6-12, 2002, Philadelphia, PA, USA.*, pages 417–424.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 347–354.