Sentiment Classification on Polarity Reviews: An Empirical Study Using Rating-based Features

Dai Quoc Nguyen[⋆] and **Dat Quoc Nguyen**[⋆] and **Thanh Vu**[†] and **Son Bao Pham**[⋆]

[⋆] Faculty of Information Technology
University of Engineering and Technology
Vietnam National University, Hanoi
{dainq, datnq, sonpb}@vnu.edu.vn

[†] Computing and Communications Department
The Open University, Milton Keynes, UK
thanh.vu@open.ac.uk

Abstract

We present a new feature type named *rating-based feature* and evaluate the contribution of this feature to the task of document-level sentiment analysis. We achieve state-of-the-art results on two publicly available standard polarity movie datasets: on the dataset consisting of 2000 reviews produced by Pang and Lee (2004) we obtain an accuracy of 91.6% while it is 89.87% evaluated on the dataset of 50000 reviews created by Maas et al. (2011). We also get a performance at 93.24% on our own dataset consisting of 233600 movie reviews, and we aim to share this dataset for further research in sentiment polarity analysis task.

1 Introduction

This paper focuses on document-level sentiment classification on polarity reviews. Specifically, the document-level sentiment analysis is to identify either a positive or negative opinion in a given opinionated review (Pang and Lee, 2008; Liu, 2010). In early work, Turney (2002) proposed an unsupervised learning algorithm to classify reviews by calculating the mutual information between a given phrase and reference words "excellent" and "poor". Pang et al. (2002) applied supervised learners of Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) to determine sentiment polarity over movie reviews. Pang and Lee (2004) presented a minimum cut-based approach to detect whether each review' sentence is more likely subjective or not. Then the sentiment of the whole document review is determined by employing a machine learning method on the document's most-subjective sentences.

Recently, most sentiment polarity classification systems (Whitelaw et al., 2005; Kennedy and Inkpen, 2006; Martineau and Finin, 2009; Maas et al., 2011; Tu et al., 2012; Wang and Manning, 2012; Nguyen et al., 2013) have obtained state-of-the-art results by employing machine learning techniques using combination of various features such as N-grams, syntactic and semantic representations as well as exploiting lexicon resources (Wilson et al., 2005; Ng et al., 2006; Baccianella et al., 2010; Taboada et al., 2011).

In this paper, we firstly introduce a novel rating-based feature for the sentiment polarity classification task. Our rating-based feature can be seen by that the scores – *which users employ to rate entities on review websites* – could bring useful information for improving the performance of classifying polarity sentiment. For a review with no associated score, we could predict a score for the review in the use of a regression model learned from an external independent dataset of reviews and their actual corresponding scores. We refer to the

Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 128–135,
Baltimore, Maryland, USA. June 27, 2014.

predicted score as the rating-based feature for learning sentiment categorization.

By combining the rating-based feature with unigrams, bigrams and trigrams, we then present the results from sentiment classification experiments on the benchmark datasets published by Pang and Lee (2004) and Maas et al. (2011).

To sum up, the contributions of our study are:

- Propose a novel rating-based feature and describe regression models learned from the external dataset to predict the feature value for the reviews in the two experimental datasets.
- Achieve state-of-the-art performances in the use of the rating-based feature for the sentiment polarity classification task on the two datasets.
- Analyze comprehensively the proficiency of the rating-based feature to the accuracy performance.
- Report additional experimental results on our own dataset containing 233600 reviews.

The paper is organized as follows: We provide some related works and describe our approach in section 2 and section 3, respectively. We detail our experiments in section 4. Finally, section 5 presents concluding remarks.

2 Related Works

Whitelaw et al. (2005) described an approach using appraisal groups such as "extremely boring", or "not really very good" for sentiment analysis, in which a semi-automatically constructed lexicon is used to return appraisal attribute values for related terms. Kennedy and Inkpen (2006) analyzed the effect of contextual valence shifters on sentiment classification of movie reviews. Martineau and Finin (2009) weighted bag-of-words in employing a delta TF-IDF function for training SVMs to classify the reviews. Maas et al. (2011) introduced a model to catch sentiment information and word meanings. Tu et al. (2012) proposed an approach utilizing high-impact parse features for convolution kernels in document-level sentiment recognition. Meanwhile, Wang and Manning (2012) obtained a strong and robust performance by identifying simple NB and SVM variants. Dahl et al. (2012) applied the restricted Boltzmann machine to learn representations capturing meaningful syntactic and semantic properties of words. In addition, Nguyen et al. (2013) constructed a two-stage sentiment classifier applying reject option, where documents rejected at the first stage are forwarded to be classified at the second stage.

3 Our Approach

We apply a supervised machine learning approach to handle the task of document-level sentiment polarity classification. For machine learning experiments, besides the N-gram features, we employ a new rating-based feature for training models.

3.1 Rating-based Feature

Our proposed rating-based feature can be seen by the fact that, on various review websites, users' reviews of entities such as products, services, events and their properties ordinarily associate to scores which the users utilize to rate the entities: a positive review mostly corresponds with a high score whereas a negative one strongly correlates to a low score. Therefore, the rated score could bring useful information to enhance the sentiment classification performance.

We consider the rated score associated to each document review as a feature named RbF for learning classification model, in which the rating-based feature RbF's value of each document review in training and test sets is estimated based on a regression model learned from an *external independent dataset* of reviews along with their actual associated scores.

3.2 N-gram Features

In most related works, unigrams are considered as the most basic features, in which each document is represented as a collection of unique unigram words where each word is considered as an individual feature.

In addition, we take into account bigrams and trigrams since a combination of unigram, bigram and trigram features (N-grams) could outperform a baseline performance based on unigram features as pointed out in (Ng et al., 2006; Martineau and Finin, 2009; Wang and Manning, 2012).

We calculate the value of the N-gram feature i^{th} by using *term frequency - inverse document frequency (tf*idf)* weighting scheme for the document D as follows:

$$Ngram_{iD} = log(1 + tf_{iD}) * log\frac{|\{D\}|}{df_i}$$

where tf_{iD} is the occurrence frequency of the feature i^{th} in document D, $|\{D\}|$ is the number of documents in the data corpus $\{D\}$, and df_i is the number of documents containing the feature i^{th}. We then normalize N-gram feature vector of the document D as follows:

$$\overrightarrow{\eta Ngram_D} = \frac{\sum_{\delta \in \{D\}} \|\overrightarrow{Ngram_\delta}\|}{|\{D\}| * \|\overrightarrow{Ngram_D}\|} * \overrightarrow{Ngram_D}$$

4 Experimental Results

4.1 Experimental Setup

Benchmark datasets. We conducted experimental evaluations on the polarity dataset PL04[1] of 2000 movie reviews constructed by Pang and Lee (2004). The dataset PL04 consists of 1000 positive and 1000 negative document reviews in which each review was split into sentences with lowercase normalization. In order to compare with other published results, we evaluate our method according to 10-fold cross-validation scheme on the dataset PL04.

In addition, we carry out experiments on a large dataset IMDB11[2] of 50000 movie reviews produced by Maas et al. (2011). The large dataset IMDB11 contains a training set of 25000 labeled reviews and a test set of 25000 labeled reviews, where training and test sets have 12500 positive reviews and 12500 negative reviews in each.

Machine learning algorithm. We utilize SVM implementation in LIBSVM[3] (Chang and Lin, 2011) for learning classification models in all our experiments on the two benchmark datasets.

Preprocess. We did not apply stop-word removal, stemming and lemmatization to the dataset in any process in our system, because such stop-words as negation words might indicate sentiment orientation, and as pointed out by Leopold and Kindermann (2002) stemming and lemmatization processes could be detrimental to accuracy.

In all experiments on PL04, *we kept 30000 most frequent N-grams in the training set for each cross-validation run over each polarity class*. After removing duplication, on an average, there are total 39950 N-gram features including 10280 unigrams, 20505 bigrams and 9165 trigrams.

On the dataset IMDB11, *it was 40000 most frequent N-grams in each polarity class* to be selected for creating feature set of 53724 N-grams consisting of 13038 unigrams, 26907 bigrams and 13779 trigrams.

RbF feature extraction procedure. We aim to create an independent dataset for learning a regression model to predict the feature RbF's value for each document review in experimental datasets. Since Maas et al. (2011) also provided 7091 IMDB movie titles[4], we used those movie titles to extract all user reviews that their associated scores[5] are not equal to either 5 or 6 from the IMDB website.

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[2] http://ai.stanford.edu/∼amaas/data/sentiment/

[3] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/. Using linear kernel, default parameter settings.

[4] http://www.imdb.com/. It is noted that the 7091 movie titles are completely different from those that were used to produce the datasets PL04 and IMDB11.

[5] The score scale ranges from 1 to 10. As the reviews corresponding to rated scores 5 or 6 are likely to be ambiguous for expressing positive or negative sentiments, we decide to ignore those 5-6 score reviews. We also abandon user reviews having no associated rated scores.

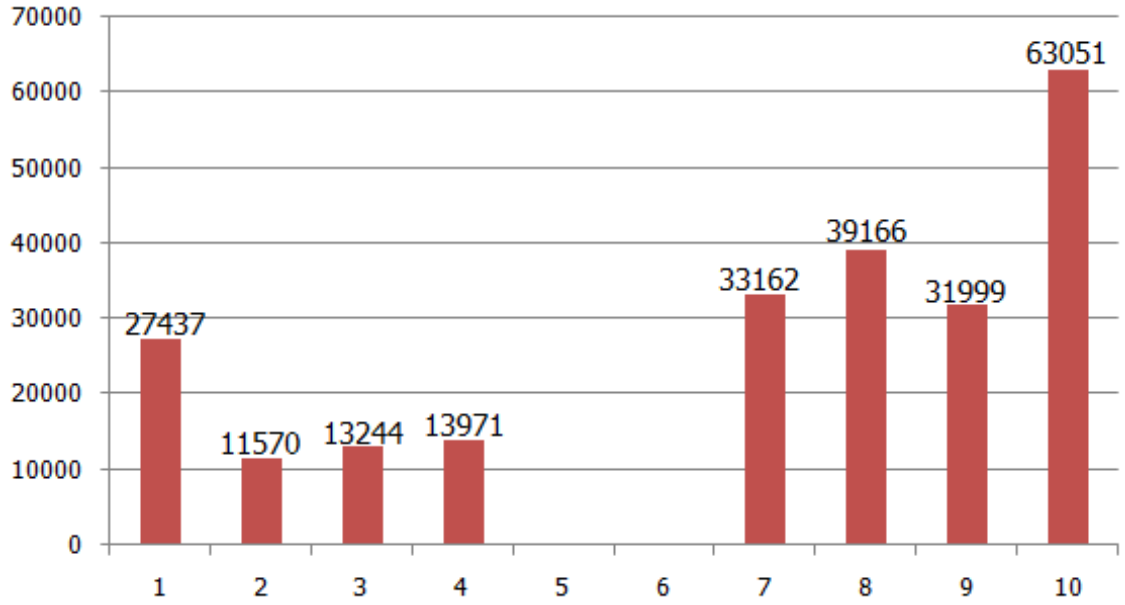

Figure 1: The score distribution of SAR14.

Consequently, we created an independent *s*core-*a*ssociated *r*eview dataset (SAR14)[6] of 233600 movie reviews and their accompanying actual scores. The external dataset SAR14 consists of 167378 user reviews connected to scores valued from 7 to 10, and 66222 reviews linked to 1-4 rated ones (as shown in Figure 1). Using SAR14, we employed Support Vector Regression algorithm implemented in SVM^{light} package[7] (Joachims, 1999) to learn the regression model employing unigram features. We then applied the learned model to predict real score values of reviews in the benchmark datasets, and referred to those values as the values of the feature RbF.

Although using N-gram features (consisting of unigrams, bigrams and trigrams) may give better results, we tend to use only unigrams for learning the regression model because of saving the training time on the large size of SAR14. Furthermore, using unigram features is good enough as presented in section 4.4. To extract the RbF feature's value for each PL04's movie review, the regression model was trained with 20000 most frequent unigrams whilst 35000 most frequent unigrams were employed to learn regression model to estimate the RbF feature for each review in the dataset IMDB11.

4.2 Results on PL04

Table 1 shows the accuracy results of our method in comparison with other state-of-the-art SVM-based performances on the dataset PL04. Our method achieves a baseline accuracy of 87.6% which is higher than baselines obtained by all other compared approaches. The accuracy based on only RbF feature is 88.2% being higher than those published in (Pang and Lee, 2004; Martineau and Finin, 2009; Nguyen et al., 2013). By exploiting a combination of unigram and RbF features, we gain a result at 89.8% which is comparable with the highest performances reached by (Whitelaw et al., 2005; Ng et al., 2006; Wang and Manning, 2012). It is evident that rising from 87.6% to 89.8% proves the effectiveness of using RbF in sentiment polarity classification.

Turning to the use of N-grams, we attain an accuracy of 89.25% which is 1.65% higher than the baseline result of 87.6%. This shows the usefulness of adding bigram and trigram

[6]The SAR14 data set is available to download at https://sites.google.com/site/nquocdai/resources

[7]http://svmlight.joachims.org/. Using with default parameter settings.

Features	PL04	IMDB11
Unigrams (baseline)	87.60	83.69
N-grams	89.25	88.67
RbF	88.20	89.14
Unigrams + RbF	89.80	84.71
N-grams + RbF	**91.60**	**89.87**
Pang and Lee (2004)	87.20	——
Whitelaw et al. (2005)	90.20	——
Ng et al. (2006)	90.50	——
Martineau and Finin (2009)	88.10	——
Maas et al. (2011)	88.90	88.89
Tu et al. (2012)	88.50	——
Dahl et al. (2012)	——	89.23
Wang and Manning (2012)	89.45	91.22
Nguyen et al. (2013)	87.95	——

Table 1: Accuracy results (in %).

features to improve the accuracy. With 91.6%, we reach a new state-of-the-art performance by combining N-gram and RbF features. We also note that our state-of-the-art accuracy is 1.1% impressively higher than the highest accuracy published by Ng et al. (2006).

4.3 Results on IMDB11

Table 1 also shows the performance results of our approach on the dataset IMDB11. Although our method gets a baseline accuracy of 83.69% which is lower than other baseline results of 88.23% and 88.29% reported by Maas et al. (2011) and Wang and Manning (2012) respectively, we achieve a noticeable accuracy of 89.14% based on only RbF feature.

Furthermore, starting at the result of 88.67% with N-gram features, we obtain a significant increase to 89.87% by employing N-gram and RbF features. Particularly, we do better than the performance at 89.23% published by Dahl et al. (2012) with a 0.64% improvement in accuracy on 160 test cases.

From our experimental results in section 4.2 and 4.3, we conclude that there are significant gains in performance results by adding bigrams and trigrams as well as RbF feature for sentiment polarity classification. Our method combining N-grams and RbF feature outperforms most other published results on the two benchmark datasets PL04 and IMDB11.

4.4 Effects of RbF to Accuracy

This section is to give a detail analysis about the effects of using RbF feature to accuracy results of our approach (as shown in Figure 2) using full combination of N-gram and RbF features in which the RbF feature is predicted by regression models learned on the dataset SAR14 in varying number **K** of most frequent unigrams from 5000 to 40000.

On the dataset PL04, the highest accuracy obtained by using only the RbF feature is 88.90% at K's value of 10000, which it is equal to that published by Maas et al. (2011). In most cases of using N-gram and RbF features, we obtain state-of-the-art results which are higher than 91%.

On the IMDB11 dataset, at K's value of 5000, we achieve the lowest accuracy of 89.29% by using N-gram and RbF features, which it is slightly higher than the accuracy of 89.23% given by Dahl et al. (2012). In cases that K's value is higher than 10000, accuracies using only RbF feature are around 89.1%, while using the full combination returns results which are higher than 89.74%.

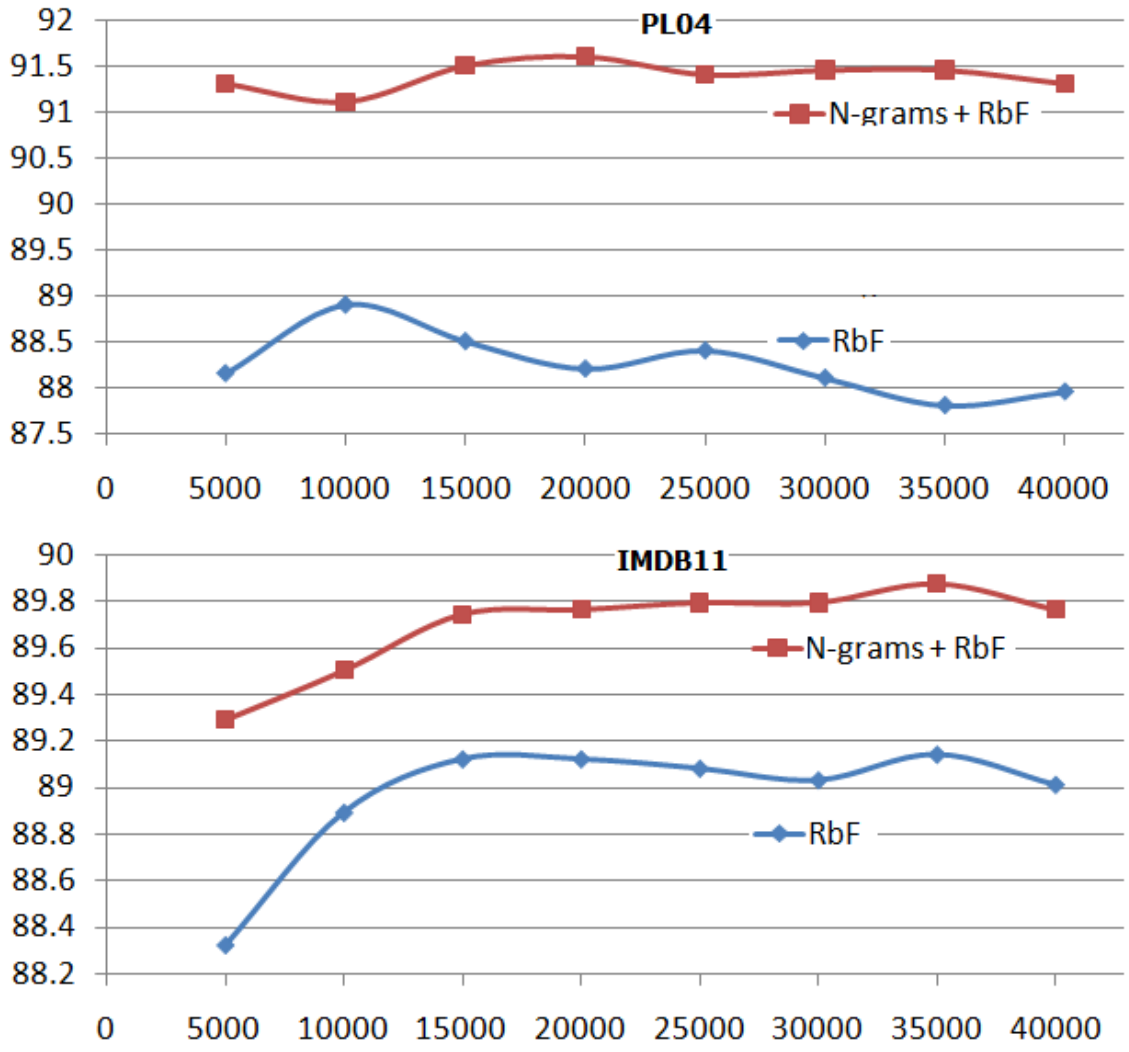

Figure 2: Effects of rating-based feature to our method's performance. The horizontal presents the number of unigram features selected for learning regression models.

4.5 Results on SAR14

As mentioned in section 4.1, our dataset SAR14 contains 233600 movie reviews. We label a review as 'positive' or 'negative' if the review has a score ≥ 7 or ≤ 4 respectively. Therefore, we create a very large dataset of 167378 positive reviews and 66222 negative reviews. Due to the large size of the dataset SAR14 and the training and classification time, we employed LIBLINEAR[8] (Fan et al., 2008) for this experiment under 10 fold cross validation scheme. We kept 50000 N-grams over each polarity class in the training set for each cross-validation run. Finally, we obtained an accuracy of *93.24%* by using N-gram features.

[8]Using L2-regularized logistic regression and setting tolerance of termination criterion to 0.01. http://www.csie.ntu.edu.tw/∼cjlin/liblinear/

5 Conclusion

In this paper, we conducted an experimental study on sentiment polarity classification. We firstly described our new rating-based feature, in which the rating-based feature is estimated based on a regression model learned from our external independent dataset SAR14 of 233600 movie reviews. We then examined the contribution of the rating-based feature and N-grams in a machine learning-based

approach on two datasets PL04 and IMDB11.

Specifically, we reach state-of-the-art accuracies at 91.6% and 89.87% on the dataset PL04 and IMDB11 respectively. Furthermore, by analyzing the effects of rating-based feature to accuracy performance, we show that the rating-based feature is very efficient to sentiment classification on polarity reviews. And adding bigram and trigram features also enhances accuracy performance. Furthermore, we get an accuracy of 93.24% on the dataset SAR14, and we also share this dataset for further research in sentiment polarity analysis task.

Acknowledgment

This work is partially supported by the Research Grant from Vietnam National University, Hanoi No. QG.14.04.

References

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 Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

George Dahl, Hugo Larochelle, and Ryan P. Adams. 2012. Training restricted boltzmann machines on word observations. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 679–686.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods: Support Vector Machines*, pages 169–184.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125.

Edda Leopold and Jörg Kindermann. 2002. Text categorization with support vector machines. how to represent texts in input space? *Mach. Learn.*, 46(1-3):423–444.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*, pages 1–38.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol 1*, pages 142–150.

Justin Martineau and Tim Finin. 2009. Delta tfidf: an improved feature space for sentiment analysis. In *Proceedings of the Third Annual Conference on Weblogs and Social Media*, pages 258–261.

Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 611–618.

Dai Quoc Nguyen, Dat Quoc Nguyen, and Son Bao Pham. 2013. A Two-Stage Classifier for Sentiment Analysis. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 897–901.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 271–278.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, pages 79–86.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June.

Zhaopeng Tu, Yifan He, Jennifer Foster, Josef van Genabith, Qun Liu, and Shouxun Lin. 2012. Identifying high-impact sub-structures for convolution kernels in document-level sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '12, pages 338–343.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '12, pages 90–94.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 625–631.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354.