CLaC-SentiPipe: SemEval2015 Subtasks 10 B,E, and Task 11

Canberk Özdemir and Sabine Bergler
CLaC Labs, Concordia University
1455 de Maisonneuve Blvd West
Montreal, Quebec, Canada, H3G 1M8
ozdemir.berkin@gmail.com, bergler@cse.concordia.ca

Abstract

CLaC Labs participated in two shared tasks for SemEval2015, Task 10 (subtasks B and E) and Task 11. The underlying system configuration is nearly identical and consists of two major components: a large Twitter lexicon compiled from tweets that carry certain selected hashtags (assumed to guarantee a sentiment polarity) and then inducing that same polarity for the words that occur in the tweets. We also use standard sentiment lexica and combine the results. The lexical sentiment features are further differentiated according to some linguistic contexts in which their triggers occur, including bigrams, negation, modality, and dependency triples. We studied feature combinations comprehensively for their interoperability and effectiveness on different datasets using the exhaustive feature combination technique of (Shareghi and Bergler, 2013a; Shareghi and Bergler, 2013b). For Subtask 10B we used a SVM, and a decision tree regressor for Task 11. The resulting systems ranked ninth for Subtask 10B, fourth for Subtask 10E, and first for Task 11.

1 Introduction

The field of Sentiment Analysis is in its second phase: initially, the task was defined, annotation standards, corpora, and feature resources were identified and provided to the research community (see (Pang and Lee, 2008)). Now, we have regular community challenges such as the SemEval Twitter Sentiment shared tasks which allow us to compare different feature choice and combination across research labs and across successive data sets. We describe here the systems we submitted to SemEval15 for Twitter Sentiment Analysis at the tweet level (Task 10B) and Figurative Language in Twitter (Task 11). The tasks and the design of the datasets is described in detail in (Rosenthal et al., 2015) for Task 10 and in (Ghosh et al., 2015) for Task 11. We also submitted a sentiment lexicon transformed from our in-house lexical resource for Task 10E.

Our system is based on a pipeline design in 5 major phases, described below. Following standard text preprocessing, we use Stanford dependencies (De Marneffe et al., 2006) and linguistic features negation, modality and their scope in connection with standard sentiment lexica from the literature and an in-house lexical resource compiled with the technique used for the NRC lexicon (Mohammad et al., 2013). These features were successful in both Task 10B (rank 9 on 40 for Twitter 2015 data, seventh on 40 for Twitter 2015 sarcasm data) and Task 11 (rank 1 of 35 runs by 15 teams). Our sentiment lexicon submitted to Task 10E ranked fourth of ten.

2 Pipeline Design

CLaCSentiPipe is a pipeline system that attempts to test the interoperability of different sentiment lexica and a selected set of linguistic annotations.

The lexical resources used are aFinn (Nielsen, 2011), MPQA (Wilson et al., 2005), BingLiu (Hu and Liu, 2004), and Gezi, our own lexical resource described below.

Third party processing resources in our GATE environment (Cunningham et al., 2013) include a hybrid of Annie and CMU tokenizers (Cunningham

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 479–485,
Denver, Colorado, June 4-5, 2015.

et al., 2002; Gimpel et al., 2011), named entity recognition (Ritter et al.,), Stanford Parser Version 3.4.1 (Socher et al., 2013) and dependency module (De Marneffe et al., 2006).

Linguistic notions used are negation and modality triggers (Kilicoglu, 2012; Rosenberg, 2013) and scope (Rosenberg, 2013) as well as dependency relations (De Marneffe et al., 2006).

Phase 1 Following tokenization, sentence splitting, POS tagging, and named entity recognition (Ritter et al.,) (to fuse multi-word names into a single token) and lookup in the sentiment lexica used, we ignore Twitter-specific items (*@name*, URLs ...) when parsing with the Stanford parser.

Phase 2 Using POS tags information for disambiguation, the prior polarity (value *positive, negative, neutral* and score where available) is determined for each token from each of the lexical resources.

Phase 3 Based on the Stanford dependencies produced in Phase 1, we identify negation and modality triggers and their scope (Rosenberg, 2013) and look up PMI scores (Church and Hanks, 1990) for dependency triples in the Gezi dependency resource.

Phase 4 The resulting features are the polarity class according to each lexical resource, embeddedness in modality or negation, as well as sentiment scores for each lexical token according to appropriate lexical resources; dependency score features using PMI scores of dependency triples and their types; dependency count features mapping PMI scores into discrete polarity classes; ad hoc features from specific annotations observed on training data.

Phase 5 The resulting feature space is grouped into subsets of features in order to create feature combinations (Shareghi and Bergler, 2013a; Shareghi and Bergler, 2013b) and processed with Weka (Witten and Frank, 2011) libSVM (Chang and Lin, 2011) with RBF kernel and parameters of cost=5, gamma=0.001 and weights=[neutral=1; positive=2; negative=2.9] for Subtask 10B and M5P (Wang and Witten, 1997), a decision tree regressor, to predict continuos values[1] for Task 11.

[1] http://www.opentox.org/dev/documentation/components/m5p

3 Lexica

In the past two years, the team that developed the NRC lexicon (Mohammad et al., 2013) dominated the Twitter sentiment task and our first question was: is the NRC lexicon itself the ultimate resource, or is the technique that derived it the essential lesson, and can that technique be reused to similar effect. We compiled a similar resource, Gezi, and compared it with the NRC lexicon, but also much smaller traditional resources, namely Bing Liu's dictionary (Hu and Liu, 2004), MPQA (Wiebe et al., 2006), and aFinn (Nielsen, 2011), a manually compiled dictionary. Extensive ablation studies showed that all the resulting dictionaries contributed to the best performing feature combination, but that the contribution of the lexica was not proportional to size (suggesting significant overlap). Surprisingly, aFinn, the smallest lexicon, by itself performs better than any of the other dictionaries by themselves and it is the one stable component in all our top performing feature combinations. In our competition system, we did not use the NRC lexicon, in order to assess whether Gezi, derived in a similar manner, was performing as well.

4 Gezi Lexical Resources

Gezi corpus To assess whether the strong performance of the NRC lexicon can be replicated and enhanced, we used their technique to compile a new resource, Gezi, by selecting positive and negative hashtags from the Twitter API from December 2013 to May 2014. The set of 35 positive and 34 negative seed hashtags were obtained from the Oxford American Writer's Thesaurus (Moody and Lindberg, 2012) by expanding the adjectives *good* and *bad*, resulting in nearly 20 million tweets, from which unigram, bigram, and dependency triple information was collected.

After removing retweets, tweets with conflicting hashtags, and tweets with little or no content words, as well as all URLs in tweets, we annotate the remaining tweets with the polarity class of their seed hashtag for our Gezi tweet corpus and project the tweet polarity onto each token inside the tweet for our unigram and bigram features.

Data processing After applying Phase 1 to the Gezi corpus the same way we use it in our main pipeline, we also parse tweets and identify negation triggers and their scopes. Then we record counts of unigrams, bigrams and dependency triples (type-head-modifier) in the context they occurred by also taking negation scope into consideration. For instance; if a term occurs in a positive-annotated tweet where it is not in the scope of a negation, its *positive* count is incremented; if it is in a positive-annotated tweet and in the scope of negation, then its *negated-positive* count is incremented. This reflects the different contexts in which the terms of the lexicon were found and associates them with the resulting sentiment. In addition, we keep terms with different POS tags separate in the resources. The counts of the terms in the *positive*, *negative*, *negated-positive* and *negated-negative* categories for the entire collection are then transformed into association scores using pointwise mutual information.

NRC and Gezi A quick comparison of Gezi and NRC unigrams and bigrams on three years of SemEval data in Table 1 shows their performance is close, with a small advantage for the much larger Gezi lexicon. Analyzing overlap of NRC (25721 unigrams) and Gezi (220399 unigrams), we find they agree only on 13957 of 16868 shared entries (both have higher agreement rates with aFinn!)

We interpret these findings as confirmation that the NRC technique can profitably be replicated and thus be used to create sentiment lexica that are bigger or smaller, that span a relevant period or contain relevant topics. We also conclude that size alone does not change results proportionally, as these large lexica clearly expand into the long tail of infrequently used words.

	SemEval Test data		
	2015	2014	2013
NRC unigrams	49.83	52.39	50.9
NRC bigrams	51.31	53.48	52.31
Gezi unigrams	54.65	60.81	57.86
Gezi bigrams	51.14	56.40	50.45
all four combined	56.07	64.26	59.60

Table 1: Comparison NRC and Gezi.

5 Features and Feature Space

Primary Features Lexicon features (*aFinn*, *NRC*, ...) encode the prior polarity of the terms in a lexicon.

Recent work in our lab on embedding predication (Kilicoglu, 2012), negation (Rosenberg, 2013), and modality (Rosenberg et al., 2012) highlighted that syntactically embedding constructions exert an influence over the meaning of constituents, so we applied this insight to sentiment values. On the 2013 dataset, most (of the 6822) tweets contained named entities (6286), as expected, but surprisingly the second most frequent feature was modality (1785), followed by negation (1356). Thus these features have the potential to influence the results to a measurable degree.

These linguistic context features were encoded as occurrences. The general schema of this integration for our system can be formulated as `polarityClass,lexicalResource,linguisticScope`, where `polarityClass` is one of *positive, negative, neutral, strong positive, strong negative*, `lexicalResource` represents a lexical resource and `linguisticScope` is one of *none, negation, modality, negation+modality*. For each tweet token, its prior polarity and any scope annotation is checked (a score feature is created if a lexicon provides score information for its terms).

The features for each feature type are aggregated into tweet-level aggregates, creating a compact feature space (94 features for Subtask 10B, 90 for Task 11).

Table 2 shows the primary features created from the aFinn lexical resource for Example 1.

(1) El Classico on a Sunday Night isn't perfect for the Monday Morning !!

This particular example has only one sentiment trigger in aFinn, *perfect*, with *aFinn score*=3 and *positive-aFinn*=1 (it is a strong positive sentiment trigger in the lexicon). In the context of Example 1, however, it occurs in the scope of a negation, thus the score is multiplied by -0.5 and the count feature *positive-aFinn-negated*=1 is activated instead, resulting in the feature assignment of Table 2.

Secondary Features The contrastive conjunction *but* and a list of contrastive adverbs (*although*, etc)

feature	value
positive-aFinn	0
positive-aFinn-negated	1
positive-aFinn-mod	0
positive-aFinn-mod-negated	0
negative-aFinn	0
negative-aFinn-negated	0
negative-aFinn-mod	0
negative-aFinn-mod-negated	0
aFinn-score	-1.5

Table 2: aFinn features for Example 1.

each constitute a feature, as do named entities. Additional ad hoc features are some special Twitter-specific POS tags (i.e. emphasis from *!!!!!*), special phrases indicative of sentiment (*can't wait*). We also found the first and last token in a tweet to carry potentially special meaning, as well as the association scores between the highest and lowest sentiment carriers in a tweet.

Feature Combinations We create feature spaces for each combination of feature subsets described above and we experiment on each combination. The submitted feature combinations for Subtask 10B and Task 11 were selected using the exhaustive feature combination technique of (Shareghi and Bergler, 2013a; Shareghi and Bergler, 2013b).

	# feat's
Primary Feature Subsets	
aFinn	9
MPQA	12
BingLiu	8
NRC unigrams	17
NRC bigrams	17
Gezi unigrams	17
Gezi bigrams	17
dependency scores	13
dependency counts	8
Secondary Feature Subsets	
pos tag based scores and counts	9
frequencies of specific annotations	12
position and top-lowest scores	6

Table 3: Feature subset bundles.

Table 3 lists the feature bundles used in our ablation studies.

6 Subtask 10B: Polarity Classification of Tweets

The task is a 3-way classification problem of labelling a tweet as *positive*, *neutral*, or *negative*, see (Rosenthal et al., 2015) for a detailed description.

We trained an SVM classifier for our experiments using last year's test sets for development. Performing manual feature selection, we selected not the feature combination that performed best on the training data but instead one that was close to the top on 2015 training data and both, 2014 and 2013 test data (for robustness) but that did not include NRC data (to better assess Gezi). The competition system included aFinn, MPQA, Bing Liu, Gezi unigrams and dependency based features in addition to all secondary features listed above.

Results The task of assigning sentiment to a tweet attracted the most participants. CLaC-SentiPipe ranked 9 of 40, a very strong placement considering less than 3% separated our results from the top ranking one. A comparison of the competing systems on the past two years' data shows that our system ranked 7 on 2013 Twitter data, 10 on 2014 Twitter data, 6 on 2014 Live Journal data, 18 on SMS messages from 2013, and 10 on Twitter 2014 Sarcasm data. This demonstrates robustness in performance. The detailed official results are shown in Table 4.

The best performing system dips to rank 12 and 13 for the LiveJournal and Sarcasm tasks of the previous years, which indicates that the different datasets compared show a certain difference, but not a big one. The very close performance of the systems in the top quarter on this task (less than 3% difference) suggests that the different approaches are drowned out by the constancy in the datasets: we may have reached the beginning of the long tail at this margin, where improvements contribute only small amounts and are not individually measurable in the general task.

7 Subtask 10E: Determining Strength of Association of Terms

SemEval 2015 Subtask 10E was a pilot task requesting association scores of terms extracted from tweets. The test set consisted of words or phrases that had to be associated with scores between 0 and

	positive			negative			neutral			overall
dataset	P	R	F1	P	R	F1	P	R	F1	F1
Twitter2015	75.58	63.20	68.84	43.51	75.34	55.17	66.63	60.08	63.19	62.00
Twitter2015-sarcasm	55.56	55.56	55.56	61.54	61.54	61.54	43.75	43.75	43.75	58.55
LiveJournal2014	79.33	66.51	72.36	68.39	82.57	74.81	67.87	68.86	68.36	73.59
SMS2013	59.26	68.29	63.46	54.39	73.86	62.65	83.55	68.60	75.34	63.05
Twitter2013	73.45	75.13	74.28	59.50	75.54	66.57	75.66	66.52	70.80	70.42
Twitter2014	78.76	70.98	74.67	58.53	74.75	65.65	63.10	66.97	64.97	70.16
Twitter2014Sarcasm	50.91	84.85	63.64	90.91	25.00	39.22	40.00	61.54	48.48	51.43

Table 4: Official CLaC-SentiPipe results for Task 10B: rank 9.

1 where 1 stands for maximum association with positive sentiment and 0 does for maximum association with negative sentiment.

We followed a simple, rule-based approach:

1. aFinn sentiment scores and Gezi (unigrams and bigrams) PMI values are used

2. if a term is part of a bigram, the unigram sentiment trigger and negation annotations are removed, if they exist

3. if a trigger is in negation scope, its prior sentiment score is multiplied with -0.5

4. if there is more than one sentiment trigger per term, the triggers' scores are summed up

5. each prior sentiment score is scaled to [0,1]

6. if there is no trigger for a term, score is 0.5

Results The evaluation metrics are Kendall and Spearman rank correlation coefficients (Nelson, 2001) for subtask 10E between gold values of words or phrases and predicted values. Gold values are human judgements from the compilation of the NRC lexicon (Kiritchenko et al., 2014).

Our simple rule-based and lexica-driven system submitted for Task 10E ranked 4th among 10 submitted systems in both correlation coefficient evaluations. Our Kendall rank correlation coefficient result is 0.584 where all results range between 0.625 and 0.254, and our Spearman rank correlation coefficient result is 0.777 where results range between 0.817 and 0.373.

8 Task 11: Figurative Language

Figurative language permeates daily life and social media, conveying non-explicit meanings using tropes such as irony, sarcasm, or metaphor. However, understanding these phenomena is not trivial for sentiment analysis systems, that usually assume that each word has only one (literal) meaning and an a priori sentiment value.

SemEval 2015 Subtask 11 Sentiment Analysis of Figurative Language in Twitter was organized for the first time this year (Ghosh et al., 2015). The challenge dataset contains tweets that contain at least one instance of figurative language and non-figurative tweets (labelled *other*). The labels are in form of sentiment scores obtained from human judgements. The dataset distinguished 3 types of figurative language, *Sarcasm, Irony* and *Metaphor*. The organizers made the tweet data available with both integer-based and float-based scores.

We tested the robustness of our linguistic embedding features by submitting the same pipeline for text processing, feature creation and the exhaustive feature combination evaluation technique of (Shareghi and Bergler, 2013a; Shareghi and Bergler, 2013b) via 10-fold cross validation on the training set with M5P (Wang and Witten, 1997), a decision tree regressor. We evaluated 10-fold cross validation predictions by calculating correlation coefficients (Nelson, 2001).

The extracted features are the same as the features we extracted for Subtask 10B. The only difference is the gold labels since Task 11 requires continuous classes while these are discrete in Subtask 10B.

We used float-based gold labels for training data and treat the problem as a regression problem. The output of our system's predictions were then

MSE				
Overall	Sarcasm	Irony	Metaphor	Other
2.117	1.023	0.779	3.155	3.411
Cosine				
Overall	Sarcasm	Irony	Metaphor	Other
0.758	0.892	0.904	0.655	0.584

Table 5: CLaC-SentiPipe in Task 11: rank 1.

rounded to integer values, as required.

Results The single submission from CLaC ranked first in both, cosine and mean squared error measures. There were wide margins between the first three systems.

The different types of figurative language were scored individually, see Table 5. In mean square error, CLaC ranked first in the *overall* score, the *metaphor*, and *other* categories. For the cosine measure, the third system of a competitor obtained best performance in the *other* category, but with a high mean squared error.

The second best system, interestingly, does not hold best performance in a single category, which demonstrates the good performance of a steady approach. The third ranked team obtained best performance for irony both in cosine similarity and least squared error, but not in their best performing (ranked) submission.

Our system has shown robustness across tasks and the linguistic features encoded have been validated for their adaptability to figurative language.

Further analysis We compared our technique with automatic forward feature selection, which interestingly selected the following six features: Gezi strong negative unigram, Gezi strong negative bigram, NRC strong positive unigram, NRC strong positive bigram: all four under scopes of both negation and modality; average scores of hashtag sentiment; counts of named entities. The results for this feature set would have been 66.41, which places it between the third and fourth-ranked systems in the competition.

This reinforces the observation that negation and modality contexts interoperate well with strong lexicon scores and are essential.

9 Conclusion

CLaCSentiPipe showed a strong top quarter performance in sentiment annotation of tweets and in its submitted lexicon, but it excelled at figurative language. We claim that the use of linguistic features negation, modality, embedding and dependency triples provides a wider context to the a priori sentiment values found in a lexicon. We combined our own large Twitter derived lexicon (Gezi) with standard resources for a range of a priori values. Gezi used the technique of extracting tweets with hashtags that are believed to guarantee sentiment polarity and inducing sentiment values for the words contained accordingly. This technique has been used for the NRC lexicon and here we showed that it can be reimplemented with good success. Our lexicon was derived from a Twitter stream of two years ago. The drastically lower performance of all systems on 2015 test data as compared to 2014 or 2013 data suggests that some events or story lines in the 2015 data use sentiment triggers differently.

Closeness of results suggest that the systems largely cover common ground, and that their specializations now fall in the area of the long tail, where incremental improvements become small and are hard to detect and measure. This confirms the observation that sentiment carrying words form a fuzzy set as demonstrated by (Andreevskaia and Bergler, 2006).

It is thus especially pleasing that the same system performed best on Task 11, sentiment annotation of tweets containing figurative language of various forms: *irony, sarcasm, metaphor*, or *other*. Here, we feel the explicit annotation of the embedding constructs has given the system the required degree of freedom to adapt to the non-literal usage. We interpret the fact that our features had not been designed specifically for this task (but were repurposed from Task 10 and merely retrained) as an indicator of robustness and a strong endorsement of our linguistically inspired features.

Acknowledgments

This work has been funded by a grant from Canada's Natural Science and Engineering Research Council (NSERC) and has benefitted from collaboration with Marc-André Faucher and Nasrin Baratalipour.

References

Alina Andreevskaia and Sabine Bergler. 2006. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the llth Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *Transactions on Intelligent Systems and Technology (TIST)*, 2(3).

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1).

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2).

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2015)*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011)*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the International Conference on Knowledge Discovery & Data Mining (KDD-2004)*.

Halil Kilicoglu. 2012. *Embedding Predications*. Ph.D. thesis, Concordia University.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval-2013)*.

Rick Moody and Christine A Lindberg. 2012. *Oxford American Writer's Thesaurus*. OUP.

Roger B. Nelson. 2001. Kendall tau metric. *Encyclopaedia of Mathematics*, 3.

Finn A. Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2).

Alan Ritter, Sam Clark, and Oren Etzion. Named entity recognition in Tweets: an experimental study.

Sabine Rosenberg, Halil Kilicoglu, and Sabine Bergler. 2012. CLaC Labs: Processing modality and negation. In *Working Notes for QA4MRE Pilot Task at CLEF 2012*.

Sabine Rosenberg. 2013. Negation triggers and their scope. Master's thesis, Concordia University.

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 (SemEval2015)*.

Ehsan Shareghi and Sabine Bergler. 2013a. CLaC-CORE: Exhaustive feature combination for measuring textual similarity. In *Proceedings of *SEM 2013 Shared Task STS*.

Ehsan Shareghi and Sabine Bergler. 2013b. Feature combination for sentence similarity. In *Proceedings of the 26th Canadian Conference on Artificial Intelligence (AI 2013)*. Springer Berlin. LNAI 7884.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Meeting of the Association for Computational Linguistics (ACL 2013)*.

Yong Wang and Ian H. Witten. 1997. Induction of model trees for predicting continuous classes. In *Poster in Proceedings of the 9th European Conference on Machine Learning*. Faculty of Informatics and Statistics, Prague.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2006. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3).

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*.

Ian H. Witten and Eibe Frank. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 3rd edition.