LT3: Sentiment Classification in User-Generated Content Using a Rich Feature Set

Cynthia Van Hee, Marjan Van de Kauter, Orphée De Clercq, Els Lefever and Véronique Hoste
LT^3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
Firstname.Lastname@UGent.be

Abstract

This paper describes our contribution to the SemEval-2014 Task 9 on sentiment analysis in Twitter. We participated in both strands of the task, viz. classification at message-level (subtask B), and polarity disambiguation of particular text spans within a message (subtask A). Our experiments with a variety of lexical and syntactic features show that our systems benefit from rich feature sets for sentiment analysis on user-generated content. Our systems ranked ninth among 27 and sixteenth among 50 submissions for task A and B respectively.

1 Introduction

Over the past few years, Web 2.0 applications such as microblogging services, social networking sites, and short messaging services have considerably increased the amount of user-generated content produced online. Millions of people rely on these services to send messages, share their views or gather information about others. Simultaneously, companies, marketeers and politicians are anxious to detect sentiment in UGC since these messages might contain valuable information about the public opinion. This explains why sentiment analysis has been a research area of great interest in the last few years (Wiebe et al., 2005; Wilson et al., 2005; Pang and Lee, 2008; Mohammad and Yang, 2011). Though first studies focussed more on product or movie reviews, we see that analyzing sentiment in UGC is currently becoming increasingly popular. The main difference between these two sources of information is that the former is rather long and contains quite formal language whereas the latter one is generally very brief and noisy and thus represents some different challenges (Maynard et al., 2012).

In this paper, we describe our contribution to the SemEval-2014 Task 9: Sentiment Analysis in Twitter (Rosenthal et al., 2014), which was a rerun of SemEval-2013 Task 2 (Nakov et al., 2013) and consisted of two subtasks:

- **Subtask A - Contextual Polarity Disambiguation:** *Given a message containing a marked instance of a word or phrase, determine whether that instance is positive, negative or neutral in that context.*

- **Subtask B - Message Polarity Classification:** *Given a message, classify whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.*

The datasets for training, development and testing were provided by the task organizers. The training datasets consisted of Twitter messages on a variety of topics. The test sets contained regular tweets (Twitter2013, Twitter2014), tweets labeled as sarcastic (TwitterSarcasm), SMS messages (SMS2013), and blog posts (LiveJournal2014). For both subtasks, the possible polarity labels were *positive*, *negative*, *neutral*, and *objective*. The datasets for subtask B contained an additional label, i.e. *objective-OR-neutral*. Table 1 presents an overview of all provided datasets. For each task and test dataset, two runs could be submitted: a constrained run using the provided training data only, and an unconstrained one using additional training data. For both tasks, we created a constrained model based on supervised learning, relying on additional lexicons and using the test datasets of SemEval-2013 as development data. Evaluation was based on averaged F-measure, considering averaged F-positive and F-negative.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 406–410,
Dublin, Ireland, August 23-24, 2014.

Dataset	Subtask A	Subtask B
Training		
Training data	26,928	9,684
Development data	1,135	1,654
Total training data	28,063	11,338
Dev-test (test SemEval-2013)		
Tweets	4,435	3,813
SMS messages	2,334	2,094
Test SemEval-2014		
Tweets + SMS messages + blog posts + sarcastic tweets	10,681	8,987

Table 1: Number of labeled instances contained by the training, development (test data SemEval-2013), and SemEval-2014 test sets.

2 System Description

Our main goal was to develop, for each polarity classification task, a classifier to label a message or an instance of that message as either positive, negative, or neutral. We ran several experiments to identify the most discriminative classifier features. This section gives an overview of the pipeline we developed and which features were implemented.

2.1 Linguistic Preprocessing

First, we performed manual cleaning on the datasets to replace non-UTF-8 characters, and we tokenized all messages using the Carnegie Mellon University Twitter Part-of-Speech Tagger (Gimpel et al., 2011). Subsequently, we Part-of-Speech tagged all instances using the CMU Twitter Part-of-Speech Tagger (Gimpel et al., 2011), and performed dependency parsing using a caseless parsing model of the Stanford parser (de Marneffe et al., 2006). Besides that, we also tagged all named entities using the Twitter NLP tools (Ritter et al., 2011) for Named Entity Recognition. As a final preprocessing step, we decided to combine the labels *neutral*, *objective* and *neutral-OR-objective*, thus recasting the task as a three-way classification task.

2.2 Feature Extraction

We implemented a number of lexical and syntactic features that represent every phrase (subtask A) or message (subtask B) within a feature vector:

N-gram features

- Word token n-gram features: a binary value for every token unigram, bigram, and trigram found in the training data.
- Character n-gram features: a binary value for every character trigram, and fourgram (within word tokens) found in the training data.
- Normalized n-gram features: n-grams that consisted of URLs and mentions or @-replies were replaced by *http://someurl* and by *@someuser*, respectively. We also normalized commonly used abbreviations [1] to their full written form (e.g. *h8* → *hate*).

Word shape features

- Character flooding: the number of word tokens with a character repeated more than two times (e.g. *soooooooo join*).
- Punctuation flooding: the number of contiguous sequences of exclamation/question marks (e.g. *GRADUATION?!?!*).
- Punctuation of the last token: a binary value indicating whether the last word token of a message contains a question/exclamation mark (e.g. *Going to Helsinki tomorrow or on the day after tomorrow, yay!*).
- The number of capitalized words (e.g. *SO EXCITED*).
- The number of hashtags (e.g. *#win*).

Lexicon features: As sentiment lexicons we consulted existing resources: AFINN (Nielsen, 2011), General Inquirer (Stone et al., 1966), MPQA (Wilson et al., 2005), NRC Emotion (Mohammad and Turney, 2010; Mohammad and Yang, 2011), Bing Liu (Hu and Liu, 2004), and Bounce (Kökciyan et al., 2013) – the latter three are Twitter-specific. Additionally, we created a list of emoticons extracted from the SemEval-2014 training data. Based on these resources, the following features were extracted:

- The number of positive, negative, and neutral lexicon words averaged over text length
- The overall polarity, which is the sum of the values of identified sentiment words

These features were extracted by 1) looking at all tokens in the instance, and 2) looking at hashtag tokens only (e.g. *win* from *#win*). We also considered negation cues by flipping the polarity

[1] These were extracted from an existing list of chat abbreviations (http://www.chatslang.com/terms/abbreviations).

sign of a sentiment word if it occurred in a negation relation (e.g. *@_2Shades_maybe 3rd team bro, he's **not better** than trey Burke from Michigan*). Negation relations were identified using the output of the dependency parser. In the example above, the positive polarity of the sentiment word *better* is flipped into negative since it occurs in a relation with *not*.

Syntactic features:

- Part-of-Speech – 25 tags, including Twitter-specific tags such as # (hashtags), @ (at-mentions), ~ (retweets), *U* (URLs or e-mail addresses), and *E* (emoticons): binary (tag occurs in the tweet or not), ternary (tag occurs zero, one, or two or more times), absolute (number of occurrences), and frequency (frequency of the tag).

- Dependency relations – four binary values for every dependency relation found in the training data. The first value indicates the presence of the lexicalized dependency relations in the test data. Additionally, as proposed by (Joshi and Penstein-Rosé, 2009), the dependency relation features are generalized in three ways: by backing off the head word to its PoS-tag, by backing off the modifier word to its PoS-tag, and by backing off both the head and modifier word.

Named entity features: This feature group consists of four features: binary (tweet contains NEs or not), absolute (number of NEs), absolute tokens (number of tokens that are part of an NE), and frequency tokens (frequency of NE tokens).

PMI features: PMI (pointwise mutual information) values indicating the association of a word with positive and negative sentiment. The higher the PMI value, the stronger the word-sentiment association. For each unigram and bigram in the training data, PMI values were extracted from the word-sentiment association lexicon created by NRC Canada (Mohammad et al., 2013). A second PMI feature was considered for each unigram based on the word-sentiment associations found in the SemEval-2014 training dataset. PMI values were calculated as follows:

$$PMI(w) = PMI(w, positive) - PMI(w, negative) \quad (1)$$

As the equation shows, the association score of a word with negative sentiment is subtracted from the word's association score with positive sentiment.

2.3 Optimizing the Classification Results

The core of our approach consisted in evaluating the aforementioned features and selecting those feature groups contributing most to the classification results. To this end, we trained an SVM classifier using the LIBSVM package (Chang and Lin, 2001) and created models for various feature combinations. A linear kernel and a cost value of 1 were chosen as parameter settings for all further experiments after cross-validation on the training data. Our experimental setup consisted of three steps: 1) training an SVM on the original training data provided by the task organizers (no development data was used), 2) generating a model, and 3) applying and evaluating the model on the development data (Twitter and SMS test data of SemEval-2013). We started our experiments with sentiment lexicon and n-gram features only, and gradually added other feature groups to identify the most contributive features. Tables 2 and 3 reveal the obtained F-scores for each step.

Features	**Dev Twitter**	**Dev SMS**
lexicons	0.6855	0.6402
n-grams	0.8482	0.8229
n-grams + lexicons	0.8628	0.8489
+ normalization n-grams	0.8632 (+ 0.0004)	0.8502 (+ 0.0013)
+ Part-of-Speech	0.8646 (+ 0.0014)	0.8582 (+ 0.0080)
+ negation	**0.8650** (+ 0.0004)	0.8654 (+ 0.0072)
+ word shape	0.8649 (- 0.0001)	0.8650 (- 0.0004)
+ named entity	0.8642 (- 0.0007)	**0.8660** (+ 0.0010)
+ dependency	0.8642 (=)	**0.8660** (=)
+ PMI	0.8610 (- 0.0032)	0.8654 (- 0.0006)

Table 2: F-scores obtained after adding other features for the Twitter and SMS development data (test data SemEval-2013) – subtask A.

Features	**Dev Twitter**	**Dev SMS**
lexicons	0.5342	0.5119
n-grams	0.5896	0.5628
n-grams + lexicons	0.6442	0.6040
+ normalization n-grams	0.6414 (- 0.0028)	0.6084 (+ 0.0044)
+ Part-of-Speech	0.6466 (+ 0.0052)	0.6333 (+ 0.0249)
+ negation	0.6542 (+ 0.0076)	0.6384 (+ 0.0051)
+ word shape	**0.6581** (+ 0.0039)	0.6394 (+ 0.0010)
+ named entity	0.6559 (- 0.0022)	0.6399 (+ 0.0005)
+ dependency	0.6467 (- 0.0092)	0.6430 (+ 0.0031)
+ PMI	0.6525 (+ 0.0058)	**0.6525** (+ 0.0095)

Table 3: F-scores obtained after adding other features for the Twitter and SMS development data (test data SemEval-2013) – subtask B.

As can be inferred from the tables, F-scores

	SMS2013	Twitter2013	LiveJournal2014	Twitter2014	Twitter2014 Sarcasm
Task A	85.26 (7/27)	86.28 (8/27)	80.44 (13/27)	81.02 (9/27)	70.76 (13/27)
Task B	64.78 (7/50)	65.56 (14/50)	68.56 (20/50)	65.47 (16/50)	47.76 (22/50)

Table 4: F-scores and rankings of our systems across the various data genres for subtask A (Contextual Polarity Disambiguation) and subtask B (Message Polarity Classification).

were already relatively high (~0.8559 for subtask A and ~0.6241 for subtask B) for the combined lexicon and n-gram features (on average 0.8559 for subtask A and 0.6241 for subtask B), which we therefore consider as a our baseline setup. Considering the results for both subtasks and data genres, we conclude that n-grams, sentiment lexicons, and PoS-tags were the most contributive feature groups, whereas named entity and dependency features did not improve the overall classification performance. However, using all feature groups (n-grams, lexicons, normalized n-grams, Part-of-Speech features, negation features, word shape features, named entity features, dependency features, and PMI features) improved the classification results (reaching an averaged F= 0.8632 for subtask A, and F= 0.6525 for subtask B) compared to classification based on lexicon (averaged F= 0.6629 for subtask A, and F= 0.5231 for subtask B) or n-gram features only (averaged F= 0.8356 for subtask A, and F= 0.5762 for subtask B). Based on these results, we conclude that using the full feature set for the classification of unseen data appears to be a promising approach, considering that it achieves good performance and that it would not tune the training model to a particular data genre.

For further optimization of the classification results, we performed feature selection in the feature groups by using a genetic algorithm approach which can explore different areas of the search space in parallel. In order to do so, we made use of the Gallop (Genetic Algorithms for Linguistic Learner Optimization) python package (Desmet et al., 2013). This enabled us to select the most contributive features from every feature group: n-gram features at token and character level, lexicon features from General Inquirer, Liu, AFINN, and Bounce, character flooding and token capitalization features, Part-of-Speech features (binary, ternary, and absolute), named entity features (binary, absolute tokens, and frequency tokens), and PMI features based on the NRC lexicon. None of the dependency relation features were selected.

3 Results

We submitted sentiment labels for the Contextual Polarity Disambiguation (subtask A) and for the Message Polarity Classification (subtask B). Our competition results are reported in Table 4. Rankings for each dataset are added between brackets. The results reveal that our systems achieved good performance in the polarity classification of unseen data across the various genres and tasks. Overall, we achieved our best classification performance on the Twitter2013 test set, obtaining an F-score of 86.28, while the best performance for this data genre is an F-score of 90.14. We saw a drop in performance on the Twitter2014 Sarcasm test set. This is consistent with most other teams as sarcastic language is hard to handle in sentiment analysis. Considering the rankings, we conclude that we performed particularly well on the SMS test dataset of SemEval-2013 for both subtasks, ranking seventh for this genre. Our systems ranked ninth among 27 submissions and sixteenth among 50 submissions for subtasks A and B respectively.

4 Conclusions and Future Work

Using a rich feature set proves to be beneficial for automatic sentiment analysis on user-generated content. Feature selection experiments revealed that features based on n-grams, sentiment lexicons, and PoS-tags were most contributive for both classification tasks, while dependency features did not contribute to overall classification performance. As future work it will be interesting to study the impact of normalization of the data on the classification performance.

Based on a shallow error analysis, we believe that including additional classification features may also be promising: modifiers other than negation cues (diminishers, increasers, modal verbs, etc.) that affect the polarity intensity, emoticon flooding, and pre- and suffixes that indicate emotion (*un-*, *dis-*, *-less*, etc.). Additionally, lemmatization and hashtag segmentation on the training data could also improve classification results.

References

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proc. of LREC'06)*.

Bart Desmet, Véronique Hoste, David Verstraeten, and Jan Verhasselt. 2013. Gallop Documentation. Technical Report LT3 13-03, University of Ghent.

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: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, KDD04*, pages 168–177, New York, NY. ACM.

Mahesh Joshi and Carolyn Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 313–316, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nadin Kökciyan, Arda Çelebi, Arzucan Özgür, and Suzan Üsküdarli. 2013. Bounce: Sentiment classification in Twitter using rich feature sets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 554–561, Atlanta, Georgia, USA. ACL.

Diane Maynard, Kalina Bontcheva, and Dominic Rout. 2012. Challenges in developing opinion mining tools for social media. In *Proc. of the LREC workshop NLP can u tag #usergeneratedcontent?!*

Saif Mohammad and Peter 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*, LA, California.

Saif Mohammad and Tony Yang. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, pages 70–79, Portland, Oregon. ACL.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *CoRR*, abs/1308.6242.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, 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, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Finn 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.*

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. *MIT Press*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Computer Intelligence*, 39(2):165–210.

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, HLT05*, pages 347–354, Stroudsburg, PA. ACL.