NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets

Xiaodan Zhu, Svetlana Kiritchenko, and Saif M. Mohammad
National Research Council Canada
Ottawa, Ontario, Canada K1A 0R6
{xiaodan.zhu,svetlana.kiritchenko,saif.mohammad}@nrc-cnrc.gc.ca

Abstract

This paper describes state-of-the-art statistical systems for automatic sentiment analysis of tweets. In a Semeval-2014 shared task (Task 9), our submissions obtained highest scores in the term-level sentiment classification subtask on both the 2013 and 2014 tweets test sets. In the message-level sentiment classification task, our submissions obtained highest scores on the LiveJournal blog posts test set, sarcastic tweets test set, and the 2013 SMS test set. These systems build on our SemEval-2013 sentiment analysis systems (Mohammad et al., 2013) which ranked first in both the term- and message-level subtasks in 2013. Key improvements over the 2013 systems are in the handling of negation. We create separate tweet-specific sentiment lexicons for terms in affirmative contexts and in negated contexts.

1 Introduction

Automatically detecting sentiment of tweets (and other microblog posts) has attracted extensive interest from both the academia and industry. The Conference on Semantic Evaluation Exercises (SemEval) organizes a shared task on the sentiment analysis of tweets with two subtasks. In the *message-level task*, the participating systems are to identify whether a tweet as a whole expresses positive, negative, or neutral sentiment. In the *term-level task*, the objective is to determine the sentiment of a marked target term (a single word or a multi-word expression) within the tweet. Our submissions stood first in both subtasks in 2013. This paper describes improvements over that system and the subsequent submissions to the 2014 shared task (Rosenthal et al., 2014).

Evaluation Set	Term-level Task	Message-level Task
Twt14	1	4
Twt13	1	2
Sarc14	3	1
LvJn14	2	1
SMS13	2	1

Table 1: Overall rank of NRC-Canada sentiment analysis models in Semeval-2014 Task 9 under the constrained condition. The rows are five evaluation datasets and the columns are the two subtasks.

The training data for the SemEval-2014 shared task is same as that of SemEval-2013 (about 10,000 tweets). The 2014 test set has five subcategories: a tweet set provided newly in 2014 (*Twt14*), the tweet set used for testing in the 2013 shared task (*Twt13*), a set of tweets that are sarcastic (*Sarc14*), a set of sentences from the blogging website LiveJournal (*LvJn14*), and the set of SMS messages used for testing in the 2013 shared task (*SMS13*). Instances from these categories were interspersed in the provided test set. The participants were not told about the source of the individual messages. The objective was to determine how well a system trained on tweets generalizes to texts from other domains.

Our submissions to SemEval-2014 Task 9, ranked first in five out of the ten subtask–dataset combinations. In the other evaluation sets as well, our submissions performed competitively. The results are summarized in Table 1. As we will show, automatically generated tweet-specific lexicons were especially helpful in all subtask–dataset combinations. The results also show that even though our models are trained only on tweets, they generalize well to data from other domains.

Our systems are based on supervised SVMs and a number of surface-form, semantic, and sentiment features. The major improvement in our 2014 system over the 2013 system is in the way it handles negation. Morante and Sporleder (2012) define negation to be "a grammatical category that allows the changing of the truth value of a proposition". Negation is often expressed through the use of negative signals or negators, words such as *isnt* and *never*, and it can significantly affect the sentiment of its scope. We create separate tweet-specific sentiment lexicons for terms in affirmative contexts and in negated contexts. That is, we automatically determine the average sentiment of a term when occurring in an affirmative context, and separately the average sentiment of a term when occurring in a negated context.

2 Our systems

Our SemEval-2014 systems are based on our SemEval-2013 systems (Mohammad et al., 2013). For completeness, we briefly revisit our previous approach, which uses support vector machine (SVM) as the classification algorithm and leverages the following features.

Lexicon features These features are generated by using three manually constructed sentiment lexicons and two automatically constructed lexicons. The manually constructed lexicons include the *NRC Emotion Lexicon* (Mohammad and Turney, 2010; Mohammad and Yang, 2011), the *MPQA Lexicon* (Wilson et al., 2005), and the *Bing Liu Lexicon* (Hu and Liu, 2004). The two automatically constructed lexicons, the *Hashtag Sentiment Lexicon* and the *Sentiment140 Lexicon*, were created specifically for tweets (Mohammad et al., 2013).

The sentiment score of each term (e.g., a word or bigram) in the automatically constructed lexicons is computed by measuring the PMI (pointwise mutual information) between the term and the positive or negative category of tweets using the formula:

$$SenScore\,(w) = PMI(w, pos) - PMI(w, neg) \quad (1)$$

where w is a term in the lexicons. $PMI(w, pos)$ is the PMI score between w and the positive class, and $PMI(w, neg)$ is the PMI score between w and the negative class. Therefore, a positive *SenScore* (w) suggests a stronger association of word w with positive sentiment and vice versa. The magnitude indicates the strength of association. Note that the sentiment class of the tweets used to construct the lexicons was automatically identified either from hashtags or from emoticons as described in (Mohammad et al., 2013).

With these lexicons available, the following features were extracted for a *text span*. Here a text span can be a target term, its context, or an entire tweet, depending on the task. The lexicon features include: (1) the number of sentiment tokens in a text span; sentiment tokens are word tokens whose sentiment scores are not zero in a lexicon; (2) the total sentiment score of the text span: $\sum_{w \in textSpan} SenScore\,(w)$; (3) the maximal score: $max_{w \in textSpan} SenScore\,(w)$; (4) the total positive and negative sentiment scores of the text span; (5) the sentiment score of the last token in the text span. Note that all these features are generated, when applicable, by using each of the sentiment lexicons mentioned above.

Ngrams We employed two types of ngram features: word ngrams and character ngrams. The former reflect the presence or absence of contiguous or non-contiguous sequences of words, and the latter are sequences of prefix/suffix characters in each word. These features are same as in our last year's submission.

Negation The number of negated contexts. Our definition of a *negated context* follows Pang et al. (2002), which will be described in more details below in Section 2.1.

POS The number of occurrences of each part-of-speech tag. We tokenized and part-of-speech tagged the tweets with the Carnegie Mellon University (CMU) Twitter NLP tool (Gimpel et al., 2011).

Cluster features The CMU POS-tagging tool provides the token clusters produced with the Brown clustering algorithm from 56 million English-language tweets. These 1,000 clusters serve as an alternative representation of tweet content, reducing the sparsity of the token space.

Encodings The encoding features are derived from hashtags, punctuation marks, emoticons, elongated words, and uppercased words.

For the term-level task, all the above features are extracted for target terms and their context, where a *context* is a window of words surrounding a target term. For the message-level task, the features are extracted from the whole tweet.

In the term-level task, we used the LIBSVM (Chang and Lin, 2011) tool with the following parameters: *-t 0 -b 1 -m 1000*. The total number of features is about 115,000. In the message-level task, we used an in-house implementation of SVM with a linear kernel. The parameter C was set to 0.005. The total number of features was about 1.5 million.

2.1 Improving lexicons and negation models

An important advantage of our SemEval-2013 systems comes from the use of the two high-coverage tweet-specific sentiment lexicons. In the SemEval-2014 submissions, we improve these lexicons by incorporating negation modeling into the lexicon generation process.

2.1.1 Improving sentiment lexicons

A word in a negated context has a different evaluative nature than the same word in an affirmative (non-negated) context. We have proposed a lexicon-based approach (Kiritchenko et al., 2014) to determining the sentiment of words in these two situations by automatically creating separate sentiment lexicons for the affirmative and negated contexts. In this way, we do not need to employ any explicit assumptions to model negation.

To achieve this, a tweet corpus is split into two parts: *Affirmative Context Corpus* and *Negated Context Corpus*. Following the work of Pang et al. (2002), we define a negated context as a segment of a tweet that starts with a negation word (e.g., *no, shouldn't*) and ends with one of the punctuation marks: ',', '.', ':', ';', '!', '?'. The list of negation words was adopted from Christopher Potts' sentiment tutorial.[1] Thus, part of a tweet that is marked as negated is included into the negated context corpus while the rest of the tweet becomes part of the affirmative context corpus. The sentiment label for the tweet is kept unchanged in both corpora. Then, we generate an affirmative context lexicon from the affirmative context corpus and a negated context lexicon from the negated context corpus using the technique described in (Kiritchenko et al., 2014).

Furthermore, we refined the method of constructing the negated context lexicons by splitting a negated context into two parts: the *immediate context* consisting of a *single* token that directly follows a negation word, and the *distant context* consisting of the rest of the tokens in the negated context. This has two benefits. Intuitively, negation affects words *directly* following the negation words more strongly than more distant words. Second, immediate-context scores are less noisy. Our simple negation scope identification algorithm can at times fail and include parts of a tweet that are not actually negated (e.g., if a punctuation mark is missing). Overall, a sentiment word can have up to three scores, one for affirmative context, one for immediate negated context, and one for distant negated context.

We reconstructed the *Hashtag Sentiment Lexicon* and the *Sentiment140 Lexicon* with this approach and used them in our SemEval-2014 systems.

2.1.2 Discriminating negation words

Different negation words, e.g., *never* and *didn't*, can have different effects on sentiment (Zhu et al., 2014; Taboada et al., 2011). In our SemEval-2014 submission, we discriminate negation words in the term-level models. For example, the word *acceptable* appearing in a sentence *this is never acceptable* is marked as *acceptable_beNever*, while in the sentence *this is not acceptable*, it is marked as *acceptable_beNot*. In this way, different negators (e.g., *be_not* and *be_never*) are treated differently. Note that we do not differentiate the *tense* and *person* of auxiliaries in order to reduce sparseness (e.g., *was not* and *am not* are treated in the same way). This new representation is used to extract ngrams and lexicon-based features.

3 Results

Overall performance The evaluation metric used in the competition is the macro-averaged F-measure calculated over the positive and negative categories. Table 2 presents the overall performance of our models. NRC13 and NRC14 are the systems we submitted to SemEval-2013 and SemEval-2014, respectively. The integers in the brackets are our official ranks in SemEval-2014 under the constrained condition.

In the term-level task, our submission ranked first on the two Tweet datasets among 14 teams. The results show that we achieved significant improvements over our last year's submission: the F-score improves from 85.19 to 86.63 on the Twt14 data and from 89.10 to 90.14 on the Twt13 data. More specifically, on the Twt14 data, the approach described in Section 2.1.1 improved our F-score

[1] http://sentiment.christopherpotts.net/lingstruc.html

	Term-level		Message-level	
	NRC13	NRC14	NRC13	NRC14
Twt14	85.19	**86.63**(1)	68.88	**69.85**(4)
Twt13	89.10	**90.14**(1)	69.02	**70.75**(2)
Sarc14	78.16	**77.13**(3)	47.64	**58.16**(1)
LvJn14	84.96	**85.49**(2)	74.01	**74.84**(1)
SMS13	88.34	**88.03**(2)	68.34	**70.28**(1)

Table 2: Overall performance of the NRC-Canada sentiment analysis systems.

from 85.19 to 86.37, and discriminating negation words (discussed in Section 2.1.2) further improved the F-score from 86.37 to 86.63.

Our system ranked second on the LvJn14 and SMS13 dataset. Note that the term-level system that ranked first on LvJn14 performed worse than our system on SMS13 and the system that ranked first on SMS13 showed worse results than ours on LvJn14, indicating that our term-level models in general have good generalizability on these two out-of-domain datasets.

On the message-level task, again the NRC14 system showed significant improvements over the last year's system on all five datasets. It achieved the second best result on the Twt13 data and the fourth result on the Twt14 data among 42 teams. It was also the best system to predict sentiment in sarcastic tweets (Sarc14). Furthermore, the system proved to generalize well to other types of short informal texts; it placed first on the two out-of-domain datasets: SMS13 and LvJn14. We observe a major improvement of our message-level model on Sarc14 over our last year's model, but as the size of Sarc14 is small (86 tweets), more data and analysis would be desirable to help better understand this phenomenon.

Contribution of features Table 3 presents the results of ablation experiments on all five test sets for the term-level task. The features derived from the manual and automatic lexicons proved to be useful on four datasets. The only exception is the Sarc14 data where removing lexicon features results in no performance improvement. Considering that this test set is very small (only about 100 test terms), further investigation would be desirable if a larger dataset becomes available. Also, in sarcasm the real sentiment of a text span may be different from its literal sentiment. In such a situation, a system that correctly recognizes the literal sentiment may actually make mistakes in capturing the real sentiment. The last two rows in Table 3 show the results obtained when the features are extracted only from the target (and not from its context) and when they are extracted only from the context of the target (and not from the target itself). Observe that even though the context may influence the polarity of the target, using target features alone is substantially more useful than using context features alone. Nonetheless, adding context features improves the F-scores in general.

On the message-level task (Table 4), the features derived from the sentiment lexicons and, in particular, from our large-coverage tweet-specific lexicons turned out to be the most influential. The use of the lexicons provided consistent gains of 9–11 percentage points not only on tweet datasets, but also on out-of-domain SMS and LiveJournal data. Note that removing the features derived from the manual lexicons as well as removing the ngram features improves the performance on the Twt14 dataset. However, this effect is not observed on the Twt13 and the out-of-domain test sets. The possible explanation of this phenomenon is minor overfitting on the tweet data.

4 Conclusions

We presented supervised statistical systems for message-level and term-level sentiment analysis of tweets. They incorporate many surface-form, semantic, and sentiment features. Among submissions from over 40 teams in the Semeval-2014 shared task "Sentiment Analysis in Twitter", our submissions ranked first in five out of the ten subtask-dataset combinations. The single most useful set of features are those obtained from automatically generated tweet-specific lexicons. We obtained significant improvements over our previous system (which ranked first in the 2013 shared task) notably by estimating the sentiment of words in affirmative and negated contexts separately. Also, since different negation words impact sentiment differently, we modeled different negation words separately in our term-level system. This too led to an improvement in F-score. The results on different kinds of evaluation sets show that even though our systems are trained only on tweets, they generalize well to text from other domains such as blog posts and SMS messages. Many of the resources we created and used are made freely available.[2]

[2]www.purl.com/net/sentimentoftweets

Experiment	Twt14	Twt13	Sarc14	LvJn14	SMS13
all features	86.63	90.14	77.13	85.49	88.03
all - lexicons	81.98	86.25	80.74	80.00	83.91
all - manu. lex.	86.08	89.25	75.32	84.13	87.69
all - auto. lex.	86.05	88.32	80.38	83.96	86.18
all - ngrams	83.31	86.67	72.95	81.58	82.41
all - target	72.93	74.19	63.09	72.21	69.34
all - context	84.40	88.83	77.22	82.99	87.97

Table 3: Term-level Task: The macro-averaged F-scores obtained on the 5 test sets with one of the feature groups removed.

Experiment	Twt14	Twt13	Sarc14	LvJn14	SMS13
all features	69.85	70.75	58.16	74.84	70.28
all - lexicons	60.59	60.04	47.17	65.80	60.56
all - manu. lex.	71.84	69.84	53.34	73.41	66.60
all - auto. lex.	63.40	65.08	47.57	71.76	66.94
all - ngrams	70.02	67.90	44.58	74.43	68.45

Table 4: Message-level Task: The macro-averaged F-scores obtained on the 5 test sets with one of the feature groups removed.

Acknowledgments

We thank Colin Cherry for providing his SVM code and for helpful discussions.

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

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 ACL*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD*, pages 168–177, New York, NY, USA. ACM.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif Mohammad. 2014. Sentiment analysis of short informal texts. *(To appear) Journal of Artificial Intelligence Research*.

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 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.

Saif M. Mohammad and Tony (Wenda) Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Portland, OR, USA.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA, June.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, Philadelphia, PA.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In *Proceedings of SemEval-2014*, Dublin, Ireland.

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of ACL*, Baltimore, Maryland, USA, June.