Kea: Sentiment Analysis of Phrases Within Short Texts

Ameeta Agrawal, Aijun An
Department of Computer Science and Engineering
York University, Toronto, Canada M3J 1P3
{ameeta, aan}@cse.yorku.ca

Abstract

Sentiment Analysis has become an increasingly important research topic. This paper describes our approach to building a system for the Sentiment Analysis in Twitter task of the SemEval-2014 evaluation. The goal is to classify a phrase within a short piece of text as positive, negative or neutral. In the evaluation, classifiers trained on Twitter data are tested on data from other domains such as SMS, blogs as well as sarcasm. The results indicate that apart from sarcasm, classifiers built for sentiment analysis of phrases from tweets can be generalized to other short text domains quite effectively. However, in cross-domain experiments, SMS data is found to generalize even better than Twitter data.

1 Introduction

In recent years, new forms of communication such as microblogging and text messaging have become quite popular. While there is no limit to the range of information conveyed by tweets and short texts, people often use these messages to share their sentiments. Working with these informal text genres presents challenges for natural language processing beyond those typically encountered when working with more traditional text genres. Tweets and short texts are shorter, the language is very informal, with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology such as, RT for "re-tweet" and #hashtags for tagging (Rosenthal et al., 2014).

Although several systems have tackled the task of analyzing sentiment from entire tweets, the task of analyzing sentiments of phrases (a word or more) within a tweet has remained largely unexplored. This paper describes the details of our system that participated in the subtask A of Semeval-2014 Task 9: Sentiment Analysis in Twitter (Rosenthal et al., 2014). The goal of this task is to determine whether a phrase within a message is positive, negative or neutral in that context. Here, a *message* indicates any short informal piece of text such as a tweet, SMS data, or a sentence from Live Journal blog, which is a social networking service where Internet users keep an online diary. A *phrase* could be a word or a few consecutive words within a message.

The novelty of this task lies in the fact that a model built using only Twitter data is used to classify instances from other short text domains such as SMS and Live Journal. Moreover, a short test corpus of sarcastic tweets is also used to test the performance of the sentiment classifier.

The main contributions of this paper include a) developing a sentiment analysis classifer for phrases; b) training on Twitter data and testing on other domains such as SMS and Live Journal data to see how well the classifier generalizes to different types of text, and c) testing on sarcastic tweets.

2 Related Work

Sentiment analysis from Twitter data has attracted much attention from the research community in the past few years (Asiaee T. et al., 2012; Go et al., 2009; Pang et al., 2002; Pang and Lee, 2004; Wilson et al., 2005). However, most of these approaches classify entire tweets by their overall sentiment (positive, negative, or neutral).

The task at hand is to classify the sentiment of a phrase within a short message. The challenges of classifying contextual polarity of phrases has been previously explored by first determining whether the phrase is neutral or polar, and then disambiguating the polarity of the polar phrases (Wilson et al., 2005). Another approach entails using

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 380–384,
Dublin, Ireland, August 23-24, 2014.

manually developed patterns (Nasukawa and Yi, 2003). Both these techniques, however, experimented with general web pages and online reviews but not Twitter data.

Previously, a few systems that participated in Semeval-2013: Sentiment Analysis in Twitter task (Wilson et al., 2013; Mohammad et al., 2013; Gunther and Furrer, 2013) tackled the problem of sentiment analysis of phrases by training on data that exclusively came from tweets and tested on a corpus made up of tweets and SMS data. This time though, the task is to see how well a system trained on tweets will perform on not only SMS data, but also blog sentences from Live Journal, as well as *sarcastic* tweets.

3 Task Setup

Formally, given a message containing a phrase (one or more words), the task is to determine whether that phrase is positive, negative or neutral in that context. We were able to download 8880 tweets (7910 for training, and 970 for development) from the corpus made available by the task organizers, where each tweet includes a phrase marked as positive, negative or neutral. Keywords and hashtags were used to identify and collect messages, which were then annotated using Amazon Mechanical Turk. This task setup is further described in the task description paper (Rosenthal et al., 2014).

The evaluation consists of Twitter data as well as surprise genres such as SMS, Live Journal and Twitter Sarcasm. The purpose of hidden test genres was to see how well a system trained on tweets will perform on previously unseen domains.

4 System Description

This section describes the system components.

4.1 Supervised Machine Learning

During development time, we experimented with various supervised machine learning classifiers, but the final model was trained using Support Vector Machines (SVM) with a linear kernel as it outperformed all other classifiers. The *c* value was empirically selected and set to 1.

4.2 Features

For all tweets, the *URL* links and *@username* mentions are replaced by "URL" and "username" placeholders, respectively. The following features were included in the final model:

- Prior polarities: Previous research (Agrawal and An, 2013; Mohammad et al., 2013) has shown prior polarities of words to be one of the most important features in contextual sentiment analysis of phrases. So, for one of the features, the sum of the sentiscores of all the terms in the phrase was computed from SentiWordNet (Esuli and Sebastiani, 2006). For another feature, the prior polarity of the phrase was estimated by averaging the positive/negative strength of all its terms by looking them up in the Subjectivity Clues database (Wilson et al., 2005).

- Emoticons: An emoticon lexicon containing frequent positive and negative emoticons, as well as some of their misspellings that are generally found in tweets, was created manually[1]. The prior positive and negative emoticon features contain the counts of all positive and negative emoticons in the phrase.

- Lengths: Counts of the total number of words in the phrase, the average number of characters in the phrase, and the total number of words in the message were included.

- Punctuation: Whether the phrase contains punctuation such as '?', '!', '...', etc.

- Clusters: Word cluster IDs were obtained for each term via unsupervised Brown clustering of tweets (Owoputi et al., 2013). For example, words such as *anyone*, *anybody*, *any1*, *ne1* and *anyonee* are all represented by cluster path *0111011110*. This allows grouping multiple (mis)spellings of a word together, which would otherwise be unique unigrams.

- Unigrams: Each phrase consists of one or more words, with the average number of words in a phrase being 2. We used only unigrams as bigrams were found to reduce the accuracy on the development set.

5 Experiments and Discussion

The task organizers made available a test data set composed of 10681 instances. Table 1 describes

[1] http://goo.gl/fh6Pjr

Test sets (# instances)	Sentiment	Example Phrase to be classified (in bold)
Twitter (6908)	positive	No school at the Cuse till Wednesday **#hyped**
	negative	i know it's in january, but i **can't wait** for Winter Jam !
	neutral	Bye bye Kyiv! See you in December **:-***
SMS (2334)	positive	later on **wanna** catch a movie?
	negative	U had ur dinner already? She just **wont believe** wat i said, haiz..
	neutral	Im free on sat ... **Ok** we watch together lor
LiveJournal (1315)	positive	And Tess you are going to **prom** too on the same day as us as well
	negative	Does not seem likely that there would be any **confusion** .
	neutral	if i am **ever** king i will make it up to you .
TwitterSarcasm (124)	positive	@ImagineMore **CHEER up.** It's Monday after all. #mondayblues
	negative	I may or may not be getting sick...perfect. **#idontwantit**
	neutral	@Ken_Rosenthal mistakes? **C'mon Kenny!! ;)**

Table 1: Test corpus details.

the breakdown of the various types of text, with example phrases that are to be classified.

As expected, Live Journal has a slightly more formal sentence structure with properly spelt words, whereas Twitter and SMS data include more creative spellings. Clearly, the sarcasm category includes messages with two contradictory sentiments in close proximity. The challenge of this task lies precisely in the fact that one classifier trained on Twitter data should be able to generalize reasonably well on different types of text.

5.1 Task Results

We participated in the *constrained* version of the task which meant working with only the provided Twitter training data without any additional annotated messages. The macro-average F1-scores of the positive and negative classes, which were the evaluation criteria for the task, of our system (trained on Twitter training data and tested on Twitter test, SMS and Live Journal blog data) are presented in Table 2.

There are two interesting observations here: firstly, even though the classifier was trained solely on tweets, it performs equally well on SMS and Live Journal data; and secondly, the sarcasm category has the poorest overall performance, unsurprisingly. This suggests that cross-domain sentiment classification of phrases in short texts is a feasible option. However, sarcasm seems to be a subtle sentiment and calls for exploring features that capture not only semantic but also syntactic nuances. The low recall of the negative sarcastic instances could be due to the fact that 30% of the negative phrases are hashtags (e.g., #don'tjudge, #smartmove, #killmenow, #sadlife, #runninglate, #asthmaticproblems, #idontwantit), that require term-splitting.

Further analysis reveals that generally the positive class has better F1-scores than the negative class across all domains, except for the SMS data. One possible reason for this could be the fact that, while in all data sets (Twitter train, Twitter test, Sarcasm test) the ratio of positive to negative instances is nearly 2:1, the SMS test set is the only one with class distribution different from the training set (with less positive instances than negative). The extremely low F1-score for the neutral class is perhaps also due to the skewed class distribution, where in all data sets, the neutral instances only make up about 4 to 9% of the data.

The positive class also has a better recall than the negative class across all domains, which suggests that the system is able to identify most of the positive test instances, perhaps due to the bigger proportion of positive training instances as well as positive words in the polarity lexicons. One simple way of improving the recall of the negative class could be by increasing the number of negative instances in the training set. In fact, in a preliminary experiment with an increased number of negative instances (resampled using SMOTE (Chawla et al., 2002)), the macro-average F1-score of the SMS data set improved by 0.5 points and that of the Sarcasm set by almost 2 points. However, there was no notable improvement in the Twitter and Live Journal test sets.

We also ran some ablation experiments on the test corpus after the submission to observe the influence of individual features on the classification

	POS.			NEG.			NEU.			AVG.
	P	R	F	P	R	F	P	R	F	
Twitter	87.6	89.7	**88.6**	82.4	76.2	79.2	23.3	28.2	25.5	**83.90**
SMS	75.9	89.9	82.3	89.8	82.4	**86.0**	32.7	10.7	16.1	**84.14**
LiveJournal	76.1	87.3	**81.3**	81.8	80.2	81.0	42.1	16.7	23.9	**81.16**
Sarcasm	77.0	93.9	**84.6**	72.2	35.1	47.3	16.7	20.0	18.2	**65.94**

Table 2: Macro-average F1-scores. P, R and F represent precision, recall and F1-score, respectively.

process. Table 3 reports the macro-average F1-scores of the experiments. The "all features*" scores here are different from those submitted as the four test corpora were tested individually here as opposed to all instances mixed into one data set. The row "- prior polarities" indicates a feature set that *excludes* the prior polarities feature, and its effect on the F1-score. MCB is the Majority Class Baseline, whereas unigrams uses only the phrase unigrams, with no additional features.

	Twitter	**SMS**	**Jour.**	**Sarc.**
MCB	39.65	31.45	33.40	39.80
unigrams	81.85	82.15	79.95	74.85
all features*	**86.20**	**87.80**	**81.90**	**78.05**
- prior polarity	***-1.8***	-0.1	-0.05	-1.95
- lengths	-0.3	0	***-0.20***	-1.3
- punctuation	-0.45	-0.45	+0.10	***-2.95***
- emoticon lex	-0.15	0	+0.05	0
- word clusters	-0.15	***-1.25***	+0.05	-0.25

Table 3: Ablation tests: Trained on Twitter only.

A few observations from the feature ablation study include:

- The prior polarities and lengths seem to be two of the most distinguishing features for Twitter and Twitter Sarcasm, whereas for SMS data, the word clusters are quite useful.
- While for Twitter Sarcasm, punctuation seems to be the most important feature, it has the opposite effect on the Live Journal blog data. This may be because the punctuation features learned from Twitter data do not translate that well to blog data due to their dissimilar writing styles.
- Even though the classifier was trained on Twitter data, it has quite a strong performance on the SMS data, which is rather unsurprising in retrospect as both genres have similar character limits, which leads to creative spellings and slang.
- While using all the features leads to almost 5 F1-score points improvement over unigrams baseline in Twitter, SMS and Sarcasm data sets, they increase only 2 F1-score points in Live Journal blog data set, suggesting that this feature set is only marginally suited for blog instances. This prompted us to explore the hypothesis: how well do SMS and Live Journal data generalize to other domains, discussed in the following section.

5.2 Cross-domain Experiments

In this section, we test how well the classifiers trained on one type of text classify other types of text. In table 4, for example, the *last row* shows the results of a model trained on Journal data (1000 instances) and tested on Twitter, SMS and Sarcasm test sets, and 10-fold cross-validated on Journal data. Since this experiment measures the generalizability of different data sets, we randomly selected 500 positive and 500 negative instances for each data set, in order to minimize the influence of the size of the training data set on the classification process. Note that this experiment does not include the neutral class. As expected, the best results on the test sets are obtained when using cross-validation (except on Twitter set). However, the model built using SMS data has the best or the second-best result overall, which suggests that out of the three types of text, it is the SMS data that generalize the best.

	Test		
	Twitter	**SMS**	**Journal**
Twitter (1000)	76.4 (cv)	80.2	78.1
SMS (1000)	**76.8**	87.1 (cv)	**79.4**
Journal (1000)	73.8	82.8	85.3 (cv)

Table 4: Cross-domain training and tests.

6 Conclusion

This paper presents the details of our system that participated in the subtask A of SemEval:2014: Sentiment Analysis in Twitter. An SVM classifier was trained on a feature set consisting of prior polarities, word clusters and various Twitter-specific features. Our experiments indicate that prior polarities are one of the most important features in the sentiment analysis of phrases from short texts. Furthermore, a classifier trained on just tweets can generalize considerably well to other texts such as SMS and blog sentences, but not to sarcasm, which calls for more research. Lastly, SMS data generalizes to other texts better than Twitter data.

Acknowledgements

We would like to thank the organizers of this task for their effort and the reviewers for their useful feedback. This research is funded in part by the Centre for Information Visualization and Data Driven Design (CIV/DDD) established by the Ontario Research Fund.

References

Ameeta Agrawal and Aijun An. 2013. Kea: Expression-level sentiment analysis from Twitter data. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, June.

Amir Asiaee T., Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. 2012. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1602–1606, New York, NY, USA. ACM.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, June.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–6.

Tobias Gunther and Lenz Furrer. 2013. Gu-mlt-lt: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, June.

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 Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, K-CAP '03, pages 70–77, New York, NY, USA. ACM.

Olutobi Owoputi, Brendan O?Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA.

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*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. Semeval-2013 task 9: Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2014)*, August.

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, Stroudsburg, PA, USA.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2013)*, June.