Kea: Expression-level Sentiment Analysis from Twitter Data

Ameeta Agrawal
Computer Science and Engineering
York University
Toronto, Canada
ameeta@cse.yorku.ca

Aijun An
Computer Science and Engineering
York University
Toronto, Canada
aan@cse.yorku.ca

Abstract

This paper describes an expression-level sentiment detection system that participated in the subtask A of SemEval-2013 Task 2: Sentiment Analysis in Twitter. Our system uses a supervised approach to learn the features from the training data to classify expressions in new tweets as positive, negative or neutral. The proposed approach helps to understand the relevant features that contribute most in this classification task.

1 Introduction

In recent years, Twitter has emerged as an ubiquitous and an opportune platform for social activity. Analyzing the sentiments of the tweets expressed by an international user-base can provide an approximate view of how people feel. One of the biggest challenges of working with tweets is their short length. Additionally, the language used in tweets is very informal, with creative spellings and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, such as, RT for "re-tweet" and #hashtags, which are a type of tagging for tweets. Although several systems tackle the task of analyzing sentiments from tweets, the task of analyzing sentiments at term or phrase-level within a tweet has remained largely unexplored.

This paper describes the details of our expression-level sentiment detection system that participated in the subtask A of SemEval-2013 Task 2: Sentiment Analysis in Twitter (Wilson et al., 2013). The goal is to mark expressions (a term or short phrases) in a tweet with their contextual polarity. This is challenging given the fact that the entire length of a tweet is restricted to just 140 characters. We describe the creation of an SVM classifier that is used to classify the contextual polarity of expressions within tweets. A feature set derived from various linguistic features, parts-of-speech tagging and prior sentiment lexicons was used to train the classifier.

2 Related Work

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

The task at hand is to classify expressions with their *contextual* sentiment. Most of these expressions can be found in sentiment lexicons already annotated with their general polarity, but the focus of this task is to detect the polarity of that expression within the context of the tweet it appears in, and therefore, given the context, the polarity of the expression might differ from that found in any lexicon. One of the primary goals of this task is to facilitate the creation of a corpus of tweets with sentiment expressions marked with their contextual sentiments.

Wilson, Wiebe and Hoffman (Wilson et al., 2005) explored the challenges of contextual polarity of sentiment expressions by first determining whether an expression is neutral or polar and then disambiguating the polarity of the polar expressions. Nasukawa and Yi (Nasukawa and Yi, 2003) classified

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 530–534, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

the polarity of target expressions using manually developed patterns. Both these approaches, however, experimented with general webpages and online reviews but not Twitter data.

3 Task Setup

This paper describes the task of recognizing contextual sentiments of expressions within a tweet. Formally, given a message containing a marked instance of a word or a phrase, the task is to determine whether that instance is positive, negative or neutral in that context.

A corpus of roughly 8343 twitter messages was made available by the task organizers, where each tweet included an expression marked as positive, negative or neutral. Also available was a development data set containing 1011 tweets with similarly marked expressions. The data sets included messages on a broad range of topics such as a mixture of entities (e.g., Gadafi, Steve Jobs), products (e.g., kindle, android phone), and events (e.g., Japan earthquake, NHL playoffs). Keywords and hashtags were used to identify and collect messages relevant to the selected topic, which were then annotated using Mechanical Turk. Further details regarding the task setup may be found in the task description paper (Wilson et al., 2013).

The evaluation consisted of classifying 4435 expressions in a Twitter data set. Furthermore, to test the generalizability of the systems, the task organizers provided a test data set consisting of 2334 SMS messages, each containing a marked expression, for which no prior training data set was made available.

4 System Description

Our aim by participating in the SemEval-2013 Sentiment Analysis in Twitter task was to investigate what features are most useful in distinguishing the different polarities. The various steps of building our system are described in detail as follows.

4.1 Tokenization

Tweets are known for being notoriously noisy due to their length restricted to just 140 characters which forces users to be creative in order to get their messages across. This poses an inherent challenge when analyzing tweets which need to undergo some significant preprocessing. The first step includes tokenizing the words in the tweet. Punctuation is identified during the tokenization process and marked for inclusion as one of the features in the feature set. This includes Twitter-specific punctuation such as "#" hashtags, specific emoticons such as ":)" and any URL links are replaced by a "URL" placeholder.

4.2 *n*-gram features

Each expression consists of one or more words, with the average number of words in an expression in the training data set found to be 2. We derive lower-case unigram and bigram as well as the full string features from the expressions which are represented by their frequency counts in the feature set. The *n*-grams were cleaned (stripped of any punctuation) before being included in the feature set as they were observed to provide better results than noisy *n*-grams. Note that the presence of punctuation did become a part of the feature set as described in 4.3. We also experimented with word-splitting, especially found in hashtags (e.g., #iamsohappy); however, contrary to our initial supposition, this step resulted in poorer results overall due to word-splitting error propagation and was therefore avoided.

4.3 POS tagging

For tagging the various parts-of-speech of a tweet, we use the POS tagger (Gimpel et al., 2011) that is especially designed to work with English data from Twitter. The tagging scheme encompasses 25 tags (please see (Gimpel et al., 2011) for the full listing), including some Twitter-specific tags (which could make up as much as 13% of all tags as shown in their annotated data set) such as "#" hashtag (indicates topic/category for tweet), "@" at-mention (indicates another user as a recipient of a tweet), "RT" re-tweets and URL or email addresses. The punctuation (such as ":-)", ":b", "(:", amongst others) from the *n*-grams is captured using the "emoticon" and "punctuation" tags that are explicitly identified by this POS tagger trained especially for tweets.

Table 1 shows an example using a subset of two POS tags for an expression (# Adj. and # Emoticon denotes the number of adjectives and emoticons respectively). Other POS tags include nouns (NN), verbs (VB) and so on. Features incorporating the information about the parts-of-speech of the expres-

Esperance will be **without star player** Youssef Msakni for the first leg of the
Champions League final against Al Ahly on Saturday. #AFRICA

Prior Polarity		Length		POS in Expression		POS in Tweet		*n*-grams			
Pos.	Neg.	Exp.	Tweet	#Adj.	#Emoticon	#Adj.	#NN	“without”	“star”	“without star”	...
0	0	3	23	0	0	1	13	1	1	1	...

Table 1: Sample feature set for an expression (denoted in bold)

sion as well as the tweet denoted by their frequencies produced better results than using a binary notation. Hence frequency counts were used in the feature set.

4.4 Prior sentiment lexicon

A prior sentiment lexicon was generated by combining four already existing polarity lexicons including the Opinion Lexicon (Hu and Liu, 2004), the SentiWordNet (Esuli and Sebastiani, 2006), the Subjectivity Clues database (Wilson et al., 2005) and the General Inquirer (Stone and Hunt, 1963). If any of the words in the expression are also found in the prior sentiment lexicon, then the frequencies of such prior positive and negative words are included as features in the feature set.

4.5 Other features

Other features found to be useful in the classification process include the length of the expression as well as the length of the tweet. A sample of the feature set is shown in Table 1.

4.6 Classifier

During development time, we experimented with different classifiers but in the end, the Support Vector Machines (SVM), using the polynomial kernel, trained over tweets from the provided train and development data outperformed all the other classifiers. The final feature set included four main features plus the *n*-grams as well as the features depicting the presence or absence of a POS in the expression and the tweet.

5 Experiments and Discussion

The task organizers made available a test data set composed of 4435 tweets where each tweet contained an instance of an expression whose sentiment was to be detected. Another test corpus of 2334 SMS messages was also used in the evaluation to test how well a system trained on tweets generalizes on other data types.

The metric for evaluating the systems is F-measure. We participated in the “constrained” version of the task which meant working with only the provided training data and no additional tweets/SMS messages or sentences with sentiment annotations were used. However, other resources such as sentiment lexicons can be incorporated into the system.

Table 2, which presents the results of our submission in this task, lists the F-score of the positive, negative and neutral classes on the Twitter test data, whereas Table 3 lists the results of the SMS message data. As it can be observed from the results, the negative sentiments are classified better than the positive ones. We reckon this may be due to the comparatively fewer ways of expressing a positive emotion, while the negative sentiment seems to have a much wider vocabulary (our sentiment lexicon has 25% less positive words than negative). Whereas the positive class has a higher precision, the negative class seems to have a more notable recall. The most striking observation, however, is the extremely low F-score for the neutral class. This may be due to the highly skewed proportion (less than 5%) of neutral instances in the training data. In future work, it will be interesting to see how balancing out the proportions of the three classes affects the classification accuracy.

We also ran some ablation experiments on the provided Twitter and SMS test data sets after the submission. Table 4 reports the findings of experiments where, for example, “- prior polarities” indicates a feature set excluding the prior polarities. The metric used here is the macro-averaged F-score of the positive and the negative class. The baseline measure implements a simple SVM classifier using only the words as unigram features in the expression. Interestingly, contrary to our hypothesis during development time, using the POS of the entire tweet was the least helpful feature. Since this was an expression level classification task, it seems that using the POS features of the entire tweet may misguide the classifier. Unsurprisingly, the prior polarities turned out to be the most important part of the feature set for this classification task as it seems that many of the expressions' contextual polarities remained same as their prior polarities.

Class	Precision	Recall	F-score
Positive	0.93	0.47	0.62
Negative	0.50	0.95	0.65
Neutral	0.15	0.12	0.13
Macro-average			**0.6394**

Table 2: Submitted results: Twitter test data

Class	Precision	Recall	F-score
Positive	0.85	0.39	0.53
Negative	0.59	0.96	0.73
Neutral	0.18	0.06	0.09
Macro-average			**0.6327**

Table 3: Submitted results: SMS test data

	Twitter	SMS
Baseline	0.821	0.824
Full feature set (submitted)	*0.639*	*0.632*
- Prior polarities	0.487	0.494
- Lengths	0.612	0.576
- POS expressions	0.646	0.615
- POS tweets	**0.855**	**0.856**

Table 4: Macro-averaged F-score results using different feature sets

6 Conclusion

This paper presented the details of our system which participated in the subtask A of SemEval-2013 Task 2: Sentiment Analysis in Twitter. An SVM classifier was trained on a feature set consiting of prior polarities, various POS and other Twitter-specific features. Our experiments indicate that prior polarities from sentiment lexicons are significant features in this expression level classification task. Furthermore, a classifier trained on just tweets can generalize considerably well on SMS message data as well. As part of our future work, we would like to explore what features are more helpful in not only classifying the positive class better, but also distinguishing neutrality from polarity.

Acknowledgments

We would like to thank the organizers of this task 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

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LRECâ06*, pages 417–422.

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.

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

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*, KDD '04, pages 168–177, New York, NY, USA. ACM.

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.

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

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. Association for Computational Linguistics.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

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.

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. Association for Computational Linguistics.

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 '13, June.