Columbia_NLP: Sentiment Detection of Sentences and Subjective Phrases in Social Media

Sara Rosenthal
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
sara@cs.columbia.edu

Apoorv Agarwal
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
apoorv@cs.columbia.edu

Kathleen McKeown
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
kathy@cs.columbia.edu

Abstract

We present two supervised sentiment detection systems which were used to compete in SemEval-2014 Task 9: Sentiment Analysis in Twitter. The first system (Rosenthal and McKeown, 2013) classifies the polarity of subjective phrases as positive, negative, or neutral. It is tailored towards online genres, specifically Twitter, through the inclusion of dictionaries developed to capture vocabulary used in online conversations (e.g., slang and emoticons) as well as stylistic features common to social media. The second system (Agarwal et al., 2011) classifies entire tweets as positive, negative, or neutral. It too includes dictionaries and stylistic features developed for social media, several of which are distinctive from those in the first system. We use both systems to participate in Subtasks A and B of SemEval-2014 Task 9: Sentiment Analysis in Twitter. We participated for the first time in Subtask B: Message-Level Sentiment Detection by combining the two systems to achieve improved results compared to either system alone.

1 Introduction

In this paper we describe two prior sentiment detection algorithms for social media. Both systems (Rosenthal and McKeown, 2013; Agarwal et al., 2011) classify the polarity of sentence phrases and tweets as positive, negative, or neutral. These algorithms were used to participate in the the expression level task (Subtask A) and message level task (Subtask B) of the SemEval-2014 Task 9: Sentiment Analysis in Twitter (Rosenthal et al., 2014) which one of the authors helped organize.

We first show improved results compared to our participation in the prior year in the expression-level task (Subtask A) by incorporating a new dictionary and new features into the system. Our focus this year was on Subtask B which we participated in for the first time. We integrated two systems to achieve improved results compared to either system alone. Our analysis shows that the first system performs better on recall while the second system performs better on precision. We used confidence metrics outputted by the systems to determine which answer should be used. This resulted in a slight improvement in the Twitter dataset compared to either system alone. In this rest of this paper, we discuss related work, the methods for each system, and experiments and results for each subtask using the data provided by Semeval-2014 Task 9: Sentiment Analysis in Twitter (Rosenthal et al., 2014).

2 Related Work

Several recent papers have explored sentiment analysis in Twitter. Go et al (2009) and Pak and Paroubek (2010) classify the sentiment of tweets containing emoticons using n-grams and POS. Barbosa and Feng (2010) detect sentiment using a polarity dictionary that includes web vocabulary and tweet-specific social media features. Bermingham and Smeaton (2010) compare polarity detection in twitter to blogs and movie reviews using lexical features.

Finally, there is a large amount of related work

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 198–202,
Dublin, Ireland, August 23-24, 2014.

through the participants of Semeval 2013 Task 2, and Semeval 2014 Task9: Sentiment Analysis in Twitter (Nakov et al., 2013; Rosenthal et al., 2014). A full list of teams and results can be found in the task description papers.

3 Phrased-Based Sentiment Detection

We developed a phrase based sentiment detection system geared towards Social Media by augmenting the state of the art system developed by Agarwal et al. (2009) to include additional dictionaries such as Wiktionary and new features such as word lengthening (e.g. hellllllloooo) and emoticons (e.g. :)) (Rosenthal and McKeown, 2013). We initially evaluated our system through our participation in the first Sentiment Analysis in Twitter task (Nakov et al., 2013). We have improved our system this year by adding a new dictionary and additional features.

3.1 Lexicons

We assign a prior polarity score to each word by using the scores provided by the Dictionary of Affect in Language (DAL) (Whissel, 1989) augmented with WordNet (Fellbaum, 1998) to improve coverage. We additionally augment it with Wiktionary, emoticon, and acronym dictionaries to improve coverage in social media (Rosenthal and McKeown, 2013). The DAL covers 50.1% of the vocabulary, 16.5% are proper nouns which we exclude due to their lack of polarity. WordNet covers 8.7% of the vocabulary and Wiktionary covers 12.5% of the vocabulary. Finally, 3.6% of the vocabulary are emoticons, acronyms, word lengthening, and forms of punctuation. 8.6% of the vocabulary is not covered which means we find a prior polarity for 96.4% of the vocabulary. In addition to these dictionaries we also use SentiWordNet (Baccianella et al., 2010) as a new distinct feature that is used in addition to the prior polarity computed from the DAL scores.

3.2 Method

We include POS tags and the top n-gram features as described in prior work (Agarwal et al., 2009; Rosenthal and McKeown, 2013). The DAL and other dictionaries are used along with a negation state machine (Agarwal et al., 2009) to determine the polarity for each word in the sentence. We include all the features described in the original system (Agarwal et al., 2009) such as the DAL scores, polar chunk n-grams, and count of syntactic chunks with their prior polarity based on the chunks position. Finally, we include several lexical-stylistic features that can occur in all datasets. We divide these features into two groups, **general**: ones that are common across online and traditional genres (e.g. exclamation points), and **social media**: one that are far more common in online genres (e.g. emoticons). The features are described in further detail in the precursor to this work (Rosenthal and McKeown, 2013). Feature selection was performed using chi-square in Weka (Hall et al., 2009).

Data Set	Majority	2013	2014
Twitter Dev	38.14	77.6	81.5
Twitter Test	42.22	N/A	76.54
Twitter Sarcasm	39.81	N/A	61.76
SMS	31.45	73.3	74.55
LiveJournal	33.42	N/A	78.19

Table 1: A comparison between the 2013 and 2014 results for Subtask A using the SemEval Twitter training corpus. All results exceed the majority baseline of the positive class significantly.

In addition we introduce some new features that were not used in the prior year. SentiWordNet (Baccianella et al., 2010) is a sentiment dictionary built upon WordNet that contains scores for each word where scores > 0 indicate the word is positive and scores < 0 indicate the word is negative. We sum the scores for each word in the phrase and use this as a single polarity feature. We found that this feature alone gave us a 2% improvement over our best results from last year. We also include some other minor features: tweet and phrase length and the position of the phrase within the tweet.

3.3 Experiments and Results

We ran all of our experiments in Weka (Hall et al., 2009) using Logistic Regression. We also experimented with other learning methods (e.g. SVM and Naive Bayes) but found that Logistic Regression worked the same or better than other methods. All results are shown using the average F-measure of the positive and negative class. The results are compared against the majority baseline of the positive class. We do not use neutral/objective as the majority class because it is not included in the average F-score in the Semeval task.

The full results in the participation of SemEval 2014: Sentiment Analysis in Twitter, Subtask A, are shown in Table 1. Our system outperforms the majority baseline significantly in all classes. Our submitted system was trained using 3-way classification (positive/negative/polarity). It included all the dictionaries from prior years and the top 100 n-grams with feature selection. In addition, it included SentiWordNet and the other new features added in 2014 which provided a 4% increase compared to our best results during the prior year (77.6% to 81.5%) and a rank of 10/20 amongst the constrained systems which used no external data. Our results on the new test set is 76.54% for a rank of 14/20. We do not do well in detecting the polarity of phrases in sarcastic tweets. This is consistent with the other teams as sarcastic tweets tend to have their polarity flipped. The improvements to our system provided a 1% boost in the SMS data with a rank of 15/20. Finally, in the LiveJournal dataset we had an F-Score of 78.19% for a rank of 12/20.

4 Message-Level Sentiment Detection

Our message-level system combines two prior systems to achieve improved results. The first system inputs an entire tweet as a "phrase" to the phrase-level sentiment detection system described in Section 3. The second system is described below.

4.1 Lexicons

The second system (Agarwal et al., 2011) makes use of two dictionaries distinctive from the other system: 1) an emoticon dictionary and 2) an acronym dictionary. The emoticon dictionary was prepared by hand-labeling 170 emoticons listed on Wikipedia.[1] For example, :) is labeled as positive whereas :=(is labeled as negative. Each emoticon is assigned a label from the following set of labels: Extremely-positive, Extremely-negative, Positive, Negative, and Neutral. We compile an acronym dictionary from an on-line resource.[2] The dictionary has translations for 5,184 acronyms. For example, lol is translated to laughing out loud.

4.2 Prior Polarity Scoring

A number of our features are based on prior polarity of words. As in the phrase-based system we too build off of prior work (Agarwal et al., 2009) by using the DAL and augmenting it with Wordnet. However, we do not follow the earlier method but use it as motivation. We consider words with with a polarity score (using the pleasantness metric from the DAL) of less than 0.5 as negative, higher than 0.8 as positive and the rest as neutral. If a word is not directly found in the dictionary, we retrieve all synonyms from Wordnet. We then look for each of the synonyms in the DAL. If any synonym is found in the DAL, we assign the original word the same pleasantness score as its synonym. If none of the synonyms is present in the DAL, the word is not associated with any prior polarity. For the given data we directly found the prior polarity of 50.1% of the words. We find the polarity of another 8.7% of the words by using WordNet. So we find prior polarity of about 58.7% of English language words.

4.3 Features

We propose a set of 50 features. We calculate these features for the whole tweet and for the last one-third of the tweet. In total, we get 100 additional features. Our features may be divided into three broad categories: ones that are primarily counts of various features and therefore the value of the feature is a natural number $\in \mathbf{N}$. Second, we include features whose value is a real number $\in \mathbf{R}$. These are primarily features that capture the score retrieved from DAL. The third category is features whose values are boolean $\in \mathbf{B}$. These are bag of words, presence of exclamation marks and capitalized text. Each of these broad categories is divided into two subcategories: Polar features and Non-polar features. We refer to a feature as polar if we calculate its prior polarity either by looking it up in DAL (extended through WordNet) or in the emoticon dictionary. All other features which are not associated with any prior polarity fall in the Non-polar category. Each of the Polar and Non-polar features is further subdivided into two categories: POS and Other. POS refers to features that capture statistics about parts-of-speech of words and Other refers to all other types of features.

A more detailed explanation of the system can be found in Agarwal et al (2011).

4.4 Combined System

Our analysis showed that the first system performs better on recall while the second system performs better on precision. We also found that there were 785 tweets in the development set where one system got it correct and the other one got it incorrect. This leaves room for a significant improvement

[1] http://en.wikipedia.org/wiki/List of emoticons

[2] http://www.noslang.com/

Experiment	Twitter			SMS	LiveJournal
	Dev	Test	Sarcasm		
Majority	29.19	34.64	27.73	19.03	27.21
Phrase-Based System	62.09	64.74	40.75	56.86	62.22
Tweet-Level System	62.4	63.73	42.41	60.54	69.44
Combined System	64.6	65.42	40.02	59.84	68.79

Table 2: A comparison between the different systems using the Twitter training corpus provided by the SemEval task for Subtask B. All results exceed the majority baseline of the positive class significantly.

compared to using each system independently. We combined the two systems for the evaluation by using the confidence provided by the phrase-based system. If the phrase-based system was $< 70\%$ confident we use the message-level system.

4.5 Experiments and Results

This task was evaluated on the Twitter dataset provided by Semeval-2013 Task 2, Subtask B. All results are shown using the average F-measure of the positive and negative class. The full results in the participation of SemEval 2014: Sentiment Analysis in Twitter, Subtask B, are shown in Table 2. All the results outperform the majority baseline of the more prominent positive polarity class significantly. The combined system outperforms the individual systems for the Twitter development and test set. It does not outperform the sarcasm test set, but this may be due to the small size; it contains only 100 tweets. The Tweet-Level system outperforms the phrase-based and combined system for the LiveJournal and SMS test sets. A closer look at the results indicated that the phrase-based system has particular difficulty with the short sentences which are more common in SMS and LiveJournal. For example, the average number of characters in a tweet is 120 whereas it is 95.6 in SMS messages (Nakov et al., 2013). Short sentences are harder because there are fewer polarity words which causes the phrase-based system to incorrectly pick neutral. In addition, short sentences are harder because the BOW feature space, which is huge and already sparse, becomes sparser and individual features start to over-fit. Part of this problem is handled by using Senti-features so the space will be less sparse.

Our ranking in the Twitter 2013 and SMS 2013 development data is 18/50 and 20/50 respectively. Our rank in the Twitter 2014 test set is 15/50 and our rank in the LiveJournal test set is 19/50. Based on our rankings it is clear that our systems are geared more towards Twitter than other social media. Finally our ranking in the Sarcasm test set is 41/50. Although this ranking is quite low, it is in fact encouraging. It indicates that the sarcasm has switched the polarity of the tweet. In the future we would like to include a system (e.g. (González-Ibáñez et al., 2011)) that can detect whether the tweet is sarcastic.

5 Discussion and Future Work

We participated in Semeval-2014 Task 9: Sentiment Analysis in Twitter Subtasks A and B. In Subtask A, we show that adding additional features related to location and using SentiWordNet gives us improvement compared to our prior system. In Subtask B, we show that combining two systems achieves slight improvements over using either system alone. Combining the two system achieves greater coverage as the systems use different emoticon and acronym dictionaries and the phrase-based system uses Wiktionary. The message-level system is geared toward entire tweets whereas the phrase-based is geared toward phrases (even though, in this case we consider the entire tweet to be a "phrase"). This is reflective in several features, such as the position of the target phrase and the syntactic chunk scores in the phrase based system and the features related to the last third of the tweet in the message-level system. In the future, we'd like to perform an error analysis to determine the source of our errors and specific examples of the kind of differences found in the two systems. Finally, we have found that at times the scores of the DAL do not line up with polarity in social media. Therefore, we would like to explore including more sentiment dictionaries instead of, or in addition to, the DAL.

6 Acknowledgements

This research was funded by the DARPA DEFT Program. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views, policies, or positions of the Department of Defense, or the U.S. Government.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 24–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon, June. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING (Posters)*, pages 36–44.

Adam Bermingham and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In Jimmy Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson, and Aijun An, editors, *CIKM*, pages 1833–1836. ACM.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

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

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

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.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Sara Rosenthal and Kathleen McKeown. 2013. Columbia nlp: Sentiment detection of subjective phrases in social media. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 478–482, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, August. The COLING 2014 Organizing Committee.

C. M. Whissel. 1989. The dictionary of affect in language. In *R. Plutchik and H. Kellerman, editors, Emotion: theory research and experience*, volume 4, London. Acad. Press.