Columbia_NLP: Sentiment Detection of Subjective Phrases in Social Media

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

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

Abstract

We present a supervised sentiment detection system that 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. We show how to incorporate these new features within a state of the art system and evaluate it on subtask A in SemEval-2013 Task 2: Sentiment Analysis in Twitter.

1 Introduction

People use social media to write openly about their personal experiences, likes and dislikes. The following sentence from Twitter is a typical example: "*Tomorrow I'm coming back from Barcelona...I don't want! :(((*". The ability to detect the sentiment expressed in social media can be useful for understanding what people think about the restaurants they visit, the political viewpoints of the day, and the products they buy. These sentiments can be used to provided targeted advertising, automatically generate reviews, and make various predictions, such as political outcomes.

In this paper we develop a sentiment detection algorithm for social media that classifies the polarity of sentence phrases as positive, negative, or neutral and test its performance in Twitter through the participation in the expression level task (subtask A) of the SemEval-2013 Task 2: Sentiment Analysis in Twitter (Wilson et al., 2013) which the authors helped organize. To do so, we build on previous work on sentiment detection algorithms for the more formal news genre, notably the work of Agarwal et al (2009), but adapt it for the language of social media, in particular Twitter. We show that exploiting lexical-stylistic features and dictionaries geared toward social media are useful in detecting sentiment.

In this rest of this paper, we discuss related work, including the state of the art sentiment system (Agarwal et al., 2009) our method is based on, the lexicons we used, our method, and experiments and results.

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. Agarwal et al (2011) perform polarity sentiment detection on the entire tweet using features that are somewhat similar to ours: the DAL, lexical features (e.g. POS and n-grams), social media features (e.g. slang and hashtags) and tree kernel features. In contrast to this related work, our approach is geared towards predicting sentiment is at the phrase level as opposed to the tweet level.

3 Lexicons

Several lexicons are used in our system. We use the DAL and expand it with WordNet, as it was used in

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 478–482, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

Corpus	DAL	NNP (Post DAL)	Word Lengthening	WordNet	Wiktionary	Emoticons	Punctuation & Numbers	Not Covered
Twitter - Train	42.9%	19.2%	1.4%	10.2%	12.7%	0.3%	1.5%	11.7%
Twitter - Dev	57.3%	13.8%	1.1%	7.1%	12.2%	0.4%	2.7%	5.4%
Twitter - Test	49.9%	15.6%	1.4%	9.6%	12.1%	0.5%	1.6%	9.3%
SMS - Test	60.1%	3.6%	0.6%	7.9%	14.7%	0.6%	1.9%	10.3%

Table 1: Coverage for each of the lexicons in the training and test corpora's.

the original work (Agarwal et al., 2009), and expand it further to use Wiktionary and an emoticon lexicon. We consider proper nouns that are not in the DAL to be objective. We also shorten words that are lengthened to see if we can find the shortened version in the lexicons (e.g. sweeeet → sweet). The coverage of the lexicons for each corpus is shown in Table 1.

3.1 DAL

The Dictionary of Affect and Language (DAL) (Whissel, 1989) is an English language dictionary of 8742 words built to measure the emotional meaning of texts. In addition to using newswire, it was also built from individual sources such as interviews on abuse, students' retelling of a story, and adolescent's descriptions of emotions. It therefore covers a broad set of words. Each word is given three scores (pleasantness - also called evaluation (ee), activeness (aa), and imagery (ii)) on a scale of 1 (low) to 3 (high). We compute the polarity of a chunk in the same manner as the original work (Agarwal et al., 2009), using the sum of the AE Space Score's ($|\sqrt{ee^2 + aa^2}|$) of each word within the chunk.

3.2 WordNet

The DAL does cover a broad set of words, but we will still often encounter words that are not included in the dictionary. Any word that is not in the DAL and is not a proper noun is accessed in WordNet (Fellbaum, 1998) [1] and, if it exists, the DAL scores of the synonyms of its first sense are used in its place. In addition to the original approach, if there are no synonyms we look at the hypernym. We then compute the average scores (ee, aa, and ii) of all the words and use that as the score for the word.

3.3 Wiktionary

We use Wiktionary, an online dictionary, to supplement the common words that are not found in WordNet and the DAL. We first examine all "form of" relationships for the word such as "doesnt" is a "misspelling of" "doesn't", and 'tonite" is an "alternate form of" "tonight". If no "form of" relationships exist, we take all the words in the definitions that have their own Wiktionary page and look up the scores for each word in the DAL. (e.g., the verb definition for *LOL* (laugh out loud) in Wiktionary is *"To laugh out loud"* with *"laugh"* having its own Wiktionary definition; it is therefore looked up in the DAL and the score for "laugh" is used for *"LOL"*.) We then compute the average scores (ee, aa, and ii) of all the words and use that as the score for the word.

3.4 Emoticon Dictionary

emoticon	:)	:D	<3	:(	;)
definition	happy	laughter	love	sad	wink

Table 2: Popular emoticons and their definitions

We created a simple lexicon to map common emoticons to a definition in the DAL. We looked at over 1000 emoticons gathered from several lists on the internet[2] and computed their frequencies within a LiveJournal blog corpus. (In the future we would like to use an external Twitter corpus). We kept the 192 emoticons that appeared at least once and mapped each emoticon to a single word definition. The top 5 emoticons and their definitions are shown in Table 2. When an emoticon is found in a tweet we look up its definition in the DAL.

4 Methods

We run our data through several pre-processing steps to preserve emoticons and expand contractions. We

[1] We cannot use SentiWordNet because we are interested in the DAL scores

[2] www.chatropolis.com, www.piology.org, en.wikipedia.org

General		Social Media	
Feature	**Example**	**Feature**	**Example**
Capital Words	Hello	Emoticons	:)
Out of Vocabulary	duh	Acronyms	LOL
Punctuation	.	Repeated Questions	???
Repeated Punctuation	#@.	Exclamation Points	!
Punctuation Count	5	Repeated Exclamations	!!!!
Question Marks	?	Word Lengthening	sweeeet
Ellipses	...	All Caps	HAHA
Avg Word Length	5	Links/Images	www.url.com

Table 3: List of lexical-stylistic features and examples.

then pre-process the sentences to add Part-of-Speech tags (POS) and chunk the sentences using the CRF tagger and chunker (Phan, 2006a; Phan, 2006b). The chunker uses three labels, 'B' (beginning), 'I' (in), and 'O' (out). The 'O' label tends to be applied to punctuation which one typically wants to ignore. However, in this context, punctuation can be very important (e.g. exclamation points, and emoticons). Therefore, we append words/phrases tagged as O to the prior B-I chunk.

We apply the dictionaries to the preprocessed sentences to generate lexical, syntactic, and stylistic features. All sets of features were reduced using chi-square in Weka (Hall et al., 2009).

4.1 Lexical and Syntactic Features

We include POS tags and the top 500 n-gram features(Agarwal et al., 2009). We experimented with different amounts of n-grams and found that more than 500 n-grams reduced performance.

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).

4.2 Lexical-Stylistic Features

We include several lexical-stylistic features (see Table 3) that can occur in all datasets. We divide these features into two groups, **general**: ones that are common across online and traditional genres, and **social media**: one that are far more common in online genres. Examples of general style features are exclamation points and ellipses. Examples of social media style features are emoticons and word lengthening. Word lengthening is a common phenomenon in social media where letters are repeated to indicate emphasis (e.g. sweeeet). It is particularly common in opinionated words (Brody and Diakopoulos, 2011). The count values of each feature was normalized by the number of words in the phrase.

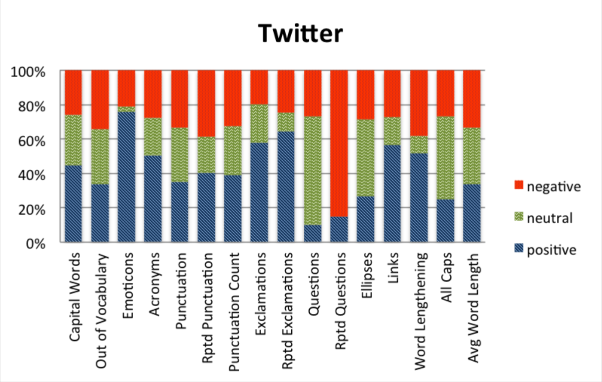

Figure 1: Percentage of lexical-stylistic features that are negative (top), neutral (middle), and positive (bottom) in the Twitter training corpus.

The percentage of lexical-stylistic features that are positive/negative/neutral is shown in Figure 1. For example, emoticons tend to indicate a positive phrase in Twitter. Each stylistic feature accounts for less than 2% of the sentence but at least one of the stylistic features exists in 61% of the Tweets.

We also computed the most frequent emoticons (*<3, :D*), acronyms (*lol*), and punctuation symbols (*#*) within a subset of the Twitter training set and included those as additional features.

5 Experiments and Results

This task was evaluated on the Twitter dataset provided by Semeval-2013 Task 2, subtask A, which the authors helped organize. Therefore, a large portion of time was spent on creating the dataset.

Experiment	Twitter		SMS
	Dev	Test	
Majority	36.3	38.1	31.5
Just DAL	70.1	72.3	67.1
WordNet	72.2	**73.6**	67.7
Wiktionary	**72.8**	**73.7**	**68.7**
Style	71.5	**73.7**	**69.7**
n-grams	**75.2**	**75.7**	**72.5**
WordNet+Style	73.2	74.6	**70.1**
Dictionaries+Style	**74.0**	**75.0**	**70.2**
Dictionaries+Style+n-grams	**75.8**	**77.6**	**73.3**

Table 4: Experiments using the Twitter corpus. Results are shown using average F-measure of the positive and negative class. All experiments include the DAL. The dictionaries refer to WordNet, Wiktionary, and Emoticon. Style refers to Lexical-Stylistic features. All results exceed the majority baseline significantly.

We ran all of our experiments in Weka (Hall et al., 2009) using Logistic Regression. We also experimented with other learning methods but found that this worked best. All results are shown using the average F-measure of the positive and negative class.

We tuned our system for Semeval-2013 Task 2, subtask A, using the provided development set and ran it on the provided Twitter and SMS test data. Our results are shown in Table 4 with all results being statistically significant over a majority baseline. We also use the DAL as a baseline to indicate how useful lexical-stylistic features (specifically those geared towards social media) and the dictionaries are in improving the performance of sentiment detection of phrases in online genres in contrast to using just the DAL. The results that are statistically significant (computed using the Wilcoxon's test, $p \leq .02$) shown in bold. Our best results for each dataset include all features with an average F-measure of 77.6% and 73.3% for the Twitter and SMS test sets respectively resulting in a significant improvement of more than 5% for each test set over the DAL baseline.

At the time of submission, we had not experimented with n-grams, and therefore chose the Dictionaries+Style system as our final version for the official run resulting in a rank of 12/22 (75% F-measure) for Twitter and 13/19 (70.2% F-measure) for SMS. Our rank with the best system, which includes n-grams, would remain the same for Twitter, but bring our rank up to 10/19 for SMS.

We looked more closely at the impact of our new features and as one would expect, feature selection found the general and social media style features (e.g. emoticons, :(, lol, word lengthening) to be useful in Twitter and SMS data. Using additional online dictionaries is useful in Twitter and SMS, which is understandable because they both have poor coverage in the DAL and WordNet. In all cases using n-grams was the most useful which indicates that context is most important. Using Dictionaries and Style in addition to n-grams did provide a significant improvement in the Twitter test set, but not in the Twitter Dev and SMS test set.

6 Conclusion and Future Work

We have explored whether social media features, Wiktionary, and emoticon dictionaries positively impact the accuracy of polarity detection in Twitter and other online genres. We found that social media related features can be used to predict sentiment in Twitter and SMS. In addition, Wiktionary helps improve the word coverage and though it does not provide a significant improvement over WordNet, it can be used in place of WordNet. On the other hand, we found that using the DAL and n-grams alone does almost as well as the best system. This is encouraging as it indicates that content is important and domain independent sentiment systems can do a good job of predicting sentiment in social media.

The results of the SMS messages dataset indicate that even though the online genres are different, the training data in one online genre can indeed be used to predict results with reasonable accuracy in the other online genre. These results show promise for further work on domain adaptation across different kinds of social media.

7 Acknowledgements

This research was partially funded by (a) the ODNI, IARPA, through the U.S. Army Research Lab and (b) 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 IARPA, the ODNI, 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.

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.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

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.

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.

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).

Xuan-Hieu Phan. 2006a. Crfchunker: Crf english phrase chunker.

Xuan-Hieu Phan. 2006b. Crftagger: Crf english phrase tagger.

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.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Alan Ritter, Sara Rosenthal, and Veselin Stoyanov. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.