SAIL: Sentiment Analysis using Semantic Similarity and Contrast Features

Nikolaos Malandrakis, Michael Falcone, Colin Vaz, Jesse Bisogni,
Alexandros Potamianos, Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA 90089, USA
{malandra,mfalcone,cvaz,jbisogni}@usc.edu,
potam@telecom.tuc.gr, shri@sipi.usc.edu

Abstract

This paper describes our submission to SemEval2014 Task 9: Sentiment Analysis in Twitter. Our model is primarily a lexicon based one, augmented by some preprocessing, including detection of Multi-Word Expressions, negation propagation and hashtag expansion and by the use of pairwise semantic similarity at the tweet level. Feature extraction is repeated for sub-strings and contrasting sub-string features are used to better capture complex phenomena like sarcasm. The resulting supervised system, using a Naive Bayes model, achieved high performance in classifying entire tweets, ranking 7th on the main set and 2nd when applied to sarcastic tweets.

1 Introduction

The analysis of the emotional content of text is relevant to numerous natural language processing (NLP), web and multi-modal dialogue applications. In recent years the increased popularity of social media and increased availability of relevant data has led to a focus of scientific efforts on the emotion expressed through social media, with Twitter being the most common subject.

Sentiment analysis in Twitter is usually performed by combining techniques used for related tasks, like word-level (Esuli and Sebastiani, 2006; Strapparava and Valitutti, 2004) and sentence-level (Turney and Littman, 2002; Turney and Littman, 2003) emotion extraction. Twitter however does present specific challenges: the breadth of possible content is virtually unlimited, the writing style is informal, the use of orthography and grammar can be "unconventional" and there are unique artifacts like hashtags. Computation systems, like those submitted to SemEval 2013 task 2 (Nakov et al., 2013) mostly use bag-of-words models with specific features added to model emotion indicators like hashtags and emoticons (Davidov et al., 2010).

This paper describes our submissions to SemEval 2014 task 9 (Rosenthal et al., 2014), which deals with sentiment analysis in twitter. The system is an expansion of our submission to the same task in 2013 (Malandrakis et al., 2013a), which used only token rating statistics as features. We expanded the system by using multiple lexica and more statistics, added steps to the pre-processing stage (including negation and multi-word expression handling), incorporated pairwise tweet-level semantic similarities as features and finally performed feature extraction on substrings and used the partial features as indicators of irony, sarcasm or humor.

2 Model Description

2.1 Preprocessing

POS-tagging / Tokenization was performed using the ARK NLP tweeter tagger (Owoputi et al., 2013), a Twitter-specific tagger.

Negations were detected using the list from Christopher Potts' tutorial. All tokens up to the next punctuation were marked as negated.

Hashtag expansion into word strings was performed using a combination of a word insertion Finite State Machine and a language model. A normalized perplexity threshold was used to detect if the output was a "proper" English string and expansion was not performed if it was not.

Multi-word Expressions (MWEs) were detected using the MIT jMWE library (Kulkarni and Finlayson, 2011). MWEs are non-compositional expressions (Sag et al., 2002), which should be

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 512–516,
Dublin, Ireland, August 23-24, 2014.

handled as a single token instead of attempting to reconstruct their meaning from their parts.

2.2 Lexicon-based features

The core of the system was formed by the lexicon-based features. We used a total of four lexica and some derivatives.

2.2.1 Third party lexica

We used three third party affective lexica.

SentiWordNet (Esuli and Sebastiani, 2006) provides continuous positive, negative and neutral ratings for each sense of every word in WordNet. We created two versions of SentiWordNet: one where ratings are averaged over all senses of a word (e.g., one ratings for "good") and one where ratings are averaged over lexeme-pos pairs (e.g., one rating for the adjective "good" and one for the noun "good").

NRC Hashtag (Mohammad et al., 2013) Sentiment Lexicon provides continuous polarity ratings for tokens, generated from a collection of tweets that had a positive or a negative word hashtag.

Sentiment140 (Mohammad et al., 2013) Lexicon provides continuous polarity ratings for tokens, generated from the sentiment140 corpus of 1.6 million tweets, with emoticons used as positive and negative labels.

2.2.2 Emotiword: expansion and adaptation

To create our own lexicon we used an automated algorithm of affective lexicon expansion based on the one presented in (Malandrakis et al., 2011; Malandrakis et al., 2013b), which in turn is an expansion of (Turney and Littman, 2002).

We assume that the continuous (in $[-1, 1]$) valence, arousal and dominance ratings of any term t_j can be represented as a linear combination of its semantic similarities d_{ij} to a set of seed words w_i and the known affective ratings of these words $v(w_i)$, as follows:

$$\hat{v}(t_j) = a_0 + \sum_{i=1}^{N} a_i \; v(w_i) \; d_{ij}, \qquad (1)$$

where a_i is the weight corresponding to seed word w_i (that is estimated as described next). For the purposes of this work, d_{ij} is the cosine similarity between context vectors computed over a corpus of 116 million web snippets (up to 1000 for each word in the Aspell spellchecker) collected using the Yahoo! search engine.

Given the starting, manually annotated, lexicon Affective Norms for English Words (Bradley and Lang, 1999) we selected 600 out of the 1034 words contained in it to serve as seed words and all 1034 words to act as the training set and used Least Squares Estimation to estimate the weights a_i. Seed word selection was performed by a simple heuristic: we want seed words to have extreme affective ratings (high absolute value) and the set to be close to balanced (sum of seed ratings equal to zero). The equation learned was used to generate ratings for any new terms.

The lexicon created by this method is task-independent, since both the starting lexicon and the raw text corpus are task-independent. To create task-specific lexica we used corpus filtering on the 116 million sentences to select ones that match our domain, using either a normalized perplexity threshold (using a maximum likelihood trigram model created from the training set tweets) or a combination of pragmatic constraints (keywords with high mutual information with the task) and perplexity threshold (Malandrakis et al., 2014). Then we re-calculated semantic similarities on the filtered corpora. In total we created three lexica: a task-independent (base) version and two adapted versions (filtered by perplexity alone and filtered by combining pragmatics and perplexity), all containing valence, arousal and dominance token ratings.

2.2.3 Statistics extraction

The lexica provide up to 17 ratings for each token. To extract tweet-level features we used simple statistics and selection criteria. First, all token unigrams and bigrams contained in a tweet were collected. Some of these n-grams were selected based on a criterion: POS tags, whether a token is (part of) a MWE, is negated or was expanded from a hashtag. The criteria were applied separately to token unigrams and token bigrams (POS tags only applied to unigrams). Then ratings statistics were extracted from the selected n-grams: length (cardinality), min, max, max amplitude, sum, average, range (max minus min), standard deviation and variance. We also created normalized versions by dividing by the same statistics calculated over all tokens, e.g., the maximum of adjectives over the maximum of all unigrams. The results of this process are features like "maximum of Emotiword valence over unigram adjectives" and "average of SentiWordNet objectivity among MWE bigrams".

2.3 Tweet-level similarity ratings

Our lexicon was formed under the assumption that semantic similarity implies affective similarity, which should apply to larger lexical units like entire tweets. To estimate semantic similarity scores between tweets we used the publicly available TakeLab semantic similarity toolkit (Šarić et al., 2012) which is based on a submission to SemEval 2012 task 6 (Agirre et al., 2012). We used the data of SemEval 2012 task 6 to train three semantic similarity models corresponding to the three datasets of that task, plus an overall model. Using these models we created four similarity ratings between each tweet of interest and each tweet in the training set. These similarity ratings were used as features of the final model.

2.4 Character features

Capitalization features are frequencies and relative frequencies at the word and letter level, extracted from all words that either start with a capital letter, have a capital letter in them (but the first letter is non-capital) or are in all capital letters.

Punctuation features are frequencies, relative frequencies and punctuation unigrams.

Character repetition features are frequencies, relative frequencies and longest string statistics of words containing a repetition of the same letter.

Emoticon features are frequencies, relative frequencies, and emoticon unigrams.

2.5 Contrast features

Cognitive Dissonance is an important phenomenon associated with complex linguistic cases like sarcasm, irony and humor (Reyes et al., 2012). To estimate it we used a simple approach, inspired by one-liner joke detection: we assumed that the final few tokens of each tweet (the "suffix") *contrast* the rest of the tweet (the "prefix") and created split versions of the tweet where the last N tokens are the suffix and all other tokens are the prefix, for $N = 2$ and $N = 3$. We repeated the feature extraction process for all features mentioned above (except for the semantic similarity features) for the prefix and suffix, nearly tripling the total number of features.

2.6 Feature selection and Training

The extraction process lead to tens of thousands of candidate features, so we performed forward stepwise feature selection using a correlation criterion (Hall, 1999) and used the resulting set of 222 features to train a model. The model chosen is a Naive Bayes tree, a tree with Naive Bayes classifiers on each leaf. The motivation comes from considering this a two stage problem: subjectivity detection and polarity classification, making a hierarchical model a natural choice. The feature selection and model training/classification was conducted using Weka (Witten and Frank, 2000).

Table 1: Performance and rank achieved by our submission for all datasets of subtasks A and B.

task	dataset	avg. F1	rank
A	LJ2014	70.62	16
	SMS2013	74.46	16
	TW2013	78.47	14
	TW2014	76.89	13
	TW2014SC	65.56	15
B	LJ2014	69.34	15
	SMS2013	56.98	24
	TW2013	66.80	10
	TW2014	67.77	7
	TW2014SC	57.26	2

Table 2: Selected features for subtask B.

Features	**number**
Lexicon-derived	178
By lexicon	
Ewrd / S140 / SWNet / NRC	71 / 53 / 33 / 21
By POS tag	
all (ignore tag)	103
adj / verb / proper noun	25 / 11 / 11
other tags	28
By function	
avg / min / sum / max	45 / 40 / 38 / 26
other functions	29
Semantic similarity	29
Punctuation	7
Emoticon	5
Other features	3
Contrast	72
prefix / suffix	54 / 18

3 Results

We took part in subtasks A and B of SemEval 2014 task 9, submitting constrained runs trained with the data the task organizers provided. Subtask B was the priority and the subtask A model was created as an afterthought: it only uses the lexicon-based and morphology features for the target string and the entire tweet as features of an NB Tree.

The overall performance of our submission on all datasets (LiveJournal, SMS, Twitter 2013, Twitter 2014 and Twitter 2014 Sarcasm) can be seen in Table 1. The subtask A system performed

Table 3: Performance on all data sets of subtask B after removing 1 set of features. Performance difference with the complete system listed if greater than 1%.

Features removed	LJ2014		SMS2013		TW2013		TW2014		TW2014SC	
	avg. F1	diff	avg. F1	diff	avg. F1	diff	avg. F1	diff	avg. F1	diff
None (Submitted)	69.3		57.0		**66.8**		67.8		57.3	
Lexicon-derived	43.6	-25.8	38.2	-18.8	49.5	-17.4	51.5	-16.3	43.5	-13.8
Emotiword	67.5	-1.9	56.4		63.5	-3.3	66.1	-1.7	54.8	-2.5
Base	68.4		56.3		65.0	-1.9	66.4	-1.4	**59.6**	2.3
Adapted	69.3		57.4		66.7		67.5		50.8	-6.5
Sentiment140	68.1	-1.3	54.5	-2.5	64.4	-2.4	64.2	-3.6	45.4	-11.9
NRC Tag	**70.6**	1.3	**58.5**	1.6	66.3		66.0	-1.7	55.3	-2.0
SentiWordNet	68.7		56.0		66.2		**68.1**		52.7	-4.6
per Lexeme	69.3		56.7		66.1		68.0		52.7	-4.5
per Lexeme-POS	68.8		57.1		66.7		67.4		55.0	-2.2
Semantic Similarity	69.0		58.2	1.2	64.9	-2.0	65.5	-2.2	52.2	-5.0
Punctuation	69.7		57.4		66.6		67.1		53.9	-3.4
Emoticon	69.3		57.0		**66.8**		67.8		57.3	
Contrast	69.2		57.5		66.7		67.0		51.9	-5.4
Prefix	69.5		57.2		**66.8**		67.2		47.4	-9.9
Suffix	68.6		57.2		66.5		67.9		56.3	

badly, ranking near the bottom (among 20 submissions) on all datasets, a result perhaps expected given the limited attention we gave to the model. The subtask B system did very well on the three Twitter datasets, ranking near the top (among 42 teams) on all three sets and placing second on the sarcastic tweets set, but did notably worse on the two non-Twitter sets.

A compact list of the features selected by the subtask B system can be seen in Table 2. The majority of features (178 of 222) are lexicon-based, 29 are semantic similarities to known tweets and the rest are mainly punctuation and emoticon features. The lexicon-based features mostly come from Emotiword, though that is probably because Emotiword contains a rating for every unigram and bigram in the tweets, unlike the other lexica. The most important part-of-speech tags are adjectives and verbs, as expected, with proper nouns being also highly important, presumably as indicators of attribution. Still, most features are calculated over all tokens (including stop words). Finally it is worth noting the 72 contrast features selected.

We also conducted a set of experiments using partial feature sets: each time we use all features minus one set, then apply feature selection and classification. The results are presented in Table 3. As expected, the lexicon-based features are the most important ones by a wide margin though the relative usefulness of the lexica changes depending on the dataset: the twitter-specific NRC lexicon actually hurts performance on non-tweets, while the task-independent Emotiword hurts performance on the sarcastic tweets set. Overall though using all is the optimal choice. Among the other features only semantic similarity provides a relatively consistent improvement.

A lot of features provide very little benefit on most sets, but virtually everything is important for the sarcasm set. Lexica, particularly the twitter specific ones like Sentiment 140 and the adapted version of Emotiword make a big difference, perhaps indicating some domain-specific aspects of sarcasm expression (though such assumptions are shaky at best due to the small size of the test set). The contrast features perform their intended function well, providing a large performance boost when dealing with sarcastic tweets and perhaps explaining our high ranking on that dataset.

Overall the subtask B system performed very well and the semantic similarity features and contrast features provide potential for further growth.

4 Conclusions

We presented a system of twitter sentiment analysis combining lexicon-based features with semantic similarity and contrast features. The system proved very successful, achieving high ranks among all competing systems in the tasks of sentiment analysis of generic and sarcastic tweets.

Future work will focus on the semantic similarity and contrast features by attempting more accurately estimate semantic similarity and using some more systematic way of identifying the "contrasting" text areas.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *proc. SemEval*, pages 385–393.

Margaret Bradley and Peter Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings. technical report C-1. The Center for Research in Psychophysiology, University of Florida.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proc. COLING*, pages 241–249.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proc. LREC*, pages 417–422.

Mark A. Hall. 1999. *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato.

Nidhi Kulkarni and Mark Alan Finlayson. 2011. jMWE: A java toolkit for detecting multi-word expressions. In *proc. Workshop on Multiword Expressions*, pages 122–124.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2011. Kernel models for affective lexicon creation. In *Proc. Interspeech*, pages 2977–2980.

Nikolaos Malandrakis, Abe Kazemzadeh, Alexandros Potamianos, and Shrikanth Narayanan. 2013a. SAIL: A hybrid approach to sentiment analysis. In *proc. SemEval*, pages 438–442.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013b. Distributional semantic models for affective text analysis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(11):2379–2392.

Nikolaos Malandrakis, Alexandros Potamianos, Kean J. Hsu, Kalina N. Babeva, Michelle C. Feng, Gerald C. Davison, and Shrikanth Narayanan. 2014. Affective language model adaptation via corpus selection. In *proc. ICASSP*, pages 4871–4874.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *proc. SemEval*, pages 321–327.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proc. SemEval*, pages 312–320.

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 *proc. NAACL*, pages 380–390.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74(0):1 – 12.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proc. SemEval*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 189–206.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proc. LREC*, volume 4, pages 1083–1086.

Peter D. Turney and Michael L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. technical report ERC-1094 (NRC 44929). National Research Council of Canada.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *proc. SemEval*, pages 441–448.

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.