UNIBA: Sentiment Analysis of English Tweets Combining Micro-blogging, Lexicon and Semantic Features

Pierpaolo Basile and **Nicole Novielli**
Department of Computer Science, University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{pierpaolo.basile,nicole.novielli}@uniba.it

Abstract

This paper describes the UNIBA team participation in the Sentiment Analysis in Twitter task (Task 10) at SemEval-2015. We propose a supervised approach relying on keyword, lexicon and micro-blogging features as well as representation of tweets in a word space.

1 Introduction

Sentiment analysis is the study of the subjectivity and polarity (positive vs. negative) of a text (Pang and Lee, 2008). With the worldwide diffusion of social media, a huge amount of textual data has been made available, thus attracting the interest of researchers in this domain (Rosenthal et al., 2014). Sentiment analysis on such informal texts poses new challenges due to the presence of slang, misspelled words and micro-blogging features such as hashtags or links and traditional approaches may not be successfully exploited in this domain. Previous research has successfully exploited approaches based on lexical and micro-blogging features (Mohammad et al., 2013). In this study, we investigate a supervised approach including three kinds of features based on keywords and micro-blogging properties of tweets, sentiment lexicons and semantics. Rather than using word-sense disambiguation (Miura et al., 2014), we represent tweets in a distributional semantic model (DSM) (Vanzo et al., 2014), which is able to learn the context of usage of words analysing cooccurrences in large corpora.

This paper describes our participation at the SemEval 2015 Sentiment Analysis in Twitter task (Rosenthal et al., 2015). We discuss methods and results of our experimental study for the overall polarity classification of tweets (*message level* subtask B). The Sentiment Analysis task focuses on English tweets. Data provided for training are annotated according to the overall polarity of each tweet (i.e., 'negative', 'positive' or 'neutral'). The system evaluation is performed on different test sets. In particular, the rank of the systems is calculated on the offical Twitter 2015 test set. Further evaluation is performed on a progress set including test instances from the previous edition of the task, to allow comparision with previous studies (Rosenthal et al., 2014). We build a supervised system based on our sentiment classifier for Italian tweets, which ranked 1st in both the polarity and subjectivity tasks at Evalita 2014 (Basile and Novielli, 2014).

The paper is structured as follows: we introduce our system and report the details about features in Section 2. We describe the evaluation and the system setup in Section 3. We conclude by reporting results and discussion in Section 4.

2 System Description

Our system is built upon our classifier for sentiment analysis of Italian tweets (Basile and Novielli, 2014). We adopt a supervised approach using Support Vector Machine as a classification algorithm. We investigate three groups of features based on: (i) keyword and micro-blogging characteristics, (ii) sentiment lexicons, and (iii) a Distributional Semantic Model (DSM).

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 595–600,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

Keywords and micro-blogging features. Keyword-based features exploit tokens occurring in the tweets (Table 1). During the tokenization we replace the user mentions, URLs and hashtags with three metatokens, "_USER_", "_URL_" and "_TAG_", for which we also count the total occurrences. As for keywords, we consider unigrams and bigrams. To deal with negations, all the n-grams occurring in a negated context receive the *neg* suffix. A negated context is a tweet fragment starting with a negation word[1] and ending with a punctuation mark (Pang et al., 2002). Moreover, we create features capturing typical aspects of micro-blogging, such as the use of upper case ratio and character repetitions[2], positive and negative emoticons, informal expressions of laughters[3], as well as the presence of exclamation and interrogative marks, negations, intensifiers [4]. Finally we include features based on word count for 1000 large-scale word clusters built on English tweets[5].

Lexicon-based Features. The second group contains features calculated for each of the eight lexicons we consider in this study. These lexicons can be differentiated based on how they represent the information about prior polarity of words.

The NRC Emotion Lexicon (Mohammad and Turney, 2010), the MPQA Lexicon (Wilson et al., 2005) and the Bing Liu Lexicon (Hu and Liu, 2004) provide lists of positive and negative words. We assign a positive score equal to 1 to the positive sentiment terms, and a negative score equal to 1 to the negative ones. Similarly, the NRC Hashtag Sentiment Lexicon and the Sentiment140 Lexicon provide a list of words with their sentiment association score, calculated as pointwise mutual information with respect to collections of positive and negative tweets (Mohammad et al., 2013). Positive and negative scores are associated, respectively, to positive and negative sentiment, while the magnitude indicates the degree of association. We consider also the lexicon used by SentiStrength[6], a state-of-the-art tool for extracting sentiment strength from informal English text on social media (Thelwall et al., 2010). The SentiStrength lexicon is structured as a list of words with scores ranging in $[-5, +5]$. A set of booster words is also provided, to increase or decrease the strength of the prior polarity of terms. Finally, we use a list of emoticons as taken from Wikipedia[7]: we assign +1 and -1 as a score for positive and negative emoticons, respectively. In all the lexicons mentioned so far either a positive or negative score is associated to each term. Using these lexicons, we extract a set of features based on prior polarity of words occuring in the tweets, as reported in Table 2. The features are computed separately for terms in affirmative contexts and terms in negated contexts.

In addition, we use SentiWordNet 3.0 (Esuli and Sebastiani, 2006). SentiWordNet extends WordNet by associating positive, negative and objective scores to each synset, where the three scores sum up to 1. A lemma can receive multiple polarity scores if it occurs in more than one synset. In such cases, we select the most frequent sense for the lemma, with respect to its part-of-speech. Thanks to the availability of the objective scores, additional features can be computed to model the presence of neutral terms, as reported in (Basile and Novielli, 2014). Also the features based on SentiWordNet are calculated separately for affirmative and negated contexts.

Finally, we consider the word classes defined in the Linguistic Inquiry and Word Count (LIWC) taxonomy, developed in the scope of psycholinguistic research (Pennebaker and Francis, 2001). LIWC organizes words into psychologically meaningful categories based on the assumption that words and language reflect most part of cognitive and emotional phenomena involved in communication. Previous research has shown how the language use varies with respect to the communicative intention, thus making possible to distinguish between objective and subjective statements as well as between agreement and disagreement expressions (Novielli and Strapparava, 2013). Therefore, we include word count features

[1] The complete list of negation words provided by Christopher Potts in his tutorial on sentiment `http://sentiment.christopherpotts.net/`.

[2] These features usually plays the same role of intensifiers in informal writing contexts.

[3] i.e., sequences of "ah".

[4] The list of booster words is the same used by Sentistrength: `http://sentistrength.wlv.ac.uk/`

[5] Twitter Word Clusters: `http://www.ark.cs.cmu.edu/TweetNLP/#resources`

[6] `http://sentistrength.wlv.ac.uk/`

[7] `http://it.wikipedia.org/wiki/Emoticon`

for each word class in LIWC. Similarly, we include word count features for the emotion word classes in the NRC Emotion Lexicon.

Semantic Features. Finally, we calculate features based on the Distributional Semantic Model (DSM). Given a set of 15M unlabelled downloaded tweets, we build a geometric space in which each word is represented as a mathematical point (Sahlgren, 2006). The similarity between words is computed as their closeness in the space. To represent a tweet in the geometric space, we adopt the superposition operator (Smolensky, 1990), that is the vector sum of all the vectors of words occurring in the tweet. We use the tweet vector $\overrightarrow{t}$ as a semantic feature in training our classifier.

In the same fashion, we build prototype vectors for each class based on the sentiment lexicons that provide prior polarity scores for words (i.e. SentiWordNet, SentiStrength, and the merge of NRC Hashtag and the Sentiment140). For example, the prototype vector for the positive class $\overrightarrow{p_{pos}}$ based on SentiStrength is obtained by summing up all the vectors of words with positive prior polarity in the SentiStrength lexicon. We use three prototype vectors to represent, for each lexicon, the positive $\overrightarrow{p_{pos}}$, negative $\overrightarrow{p_{neg}}$, and subjective $\overrightarrow{p_s}$ class (defined by considering both positive and negative words). In the case of SentiWordNet, objectivity scores are also available and allow us to build a prototype for objectivity $\overrightarrow{p_o}$. To capture the subjectivity and the polarity of a tweet $\overrightarrow{t}$, we compute the cosine similarity between $\overrightarrow{t}$ and each prototype vector.

3 Evaluation

The message level subtask (subtask B) is designed for evaluating systems on their ability to predict the overall polarity of a given tweet, with respect to three classes: positive, negative, and neutral.

Organizers provided 8,006 manually annotated tweets as training data. We use the training set[8] to extract the features described in Section 2. Details on our system setup are reported in Section 3.1. As test set, organizers provided a collection of 2,390 manually annotated tweets (Official 2015 test set). Further data from different sources (8,987 tweets overall) are included in the progress test set and are provided to allow comparison with systems participating in previous editions. Systems are compared against the gold standard of the official test set in terms of macro average F measure calculated over the positive and negative classes. For the sake of completeness, we report also weighted F measure considering all the three categories in the classification task (see Section 4).

3.1 System Setup

The system is completely developed in JAVA. We used the Liblinear[9] implementation of L_2-loss support vector classifier. Tweets are tokenized using the Twitter NLP and Part-of-Speech Tagging API[10]. We use both the tokenizer and the part-of-speech tagger to preprocess the data.

Regarding the DSM, we download 15 million tweets using the Twitter Streaming API. Tweets are downloaded by querying the API using three lexicons extracted from the training data for each class, based on Kullback-Leibler divergence (KLD) as described in (Basile and Novielli, 2014).

We download the same number of tweets for each lexicon. We exploit these unlabeled tweets to build a DSM, using the “word2vec”[11] tool based on a revised implementation of the Recurrent Neural Net Language Model (Mikolov et al., 2013) using a log-linear approach. We use the skipgram model, which is more accurate in presence of infrequent words, with 300 vector dimensions and remove the terms with less than ten occurrences, obtaining 308,493 terms overall.

In training our classifier, we set the C parameter to 0.01. We select this value after a 10-fold validation on training data to select the best combination. The total number of features exploited is 145,967.

4 Results and Discussion

The final ranking issued by the organizers considers the system performance in terms of average between F measures for the positive and negative classes only. Table 3 reports the system performance and

[8] Further development data provided by the organizers are not used for training

[9] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[10] http://www.ark.cs.cmu.edu/TweetNLP/

[11] https://code.google.com/p/word2vec/

Keyword and micro-blogging features	
$n-grams$	uni- and bi-grams are considered. User mentions, URLs and hashtag are replaced with metatokens
$count_{USER}$	total occurrences of user mentions
$count_{URL}$	total occurrences of URLs
$count_{TAG}$	total occurrences of hashtags
$uppercase_{ratio}$	the ratio between the number of upper case characters and the total number of characters
emo_{pos}	the number of positive emoticons
emo_{neg}	the number of negative emoticons
$count_{Laugh}$	the count of sequences of 'ah' as slang expression of laughters
$count_{Intensif}$	the ratio between the number of tokens with repeated characters and the total number of tokens
$count_{QMark}$	the total occurrences of question marks
$count_{ExMark}$	the total occurrences of exclamation marks
$count_{Negation}$	the total occurrences of negation words
$count_{cluster_i}$	the total occurrences of words belonging to the *i*-th cluster

Table 1: Description of keyword and micro-blogging features.

Sentiment lexicon based features	
o_{pos}	the number of tokens with positive score
o_{neg}	the number of tokens with negative score
o_{subj}	the number of tokens with either positive or negative score
$last_{pos}$	the score of the last positive token in the tweet
$last_{neg}$	the score of the last negative token in the tweet
$last_{emo}$	the score of the last emoticon in the tweet
sum_{pos}	the sum of positive scores for the tokens in the tweet
sum_{neg}	the sum of negative scores for the tokens in the tweet
sum_{subj}	the subjectivity polarity, it is the sum of the positive and negative scores
$sum_{Max_{pos}}$	the maximum positive score observed for tokens in the tweet
$sum_{Max_{neg}}$	the maximum negative score observed for tokens in the tweet
$count_{C_i}$	the total occurrences of words belonging to the i-th word class C_i, where word classes are defined by the LIWC and NRC Emotion Lexicon taxonomies

Table 2: Description of sentiment lexicon features.

its rank. The system rank on the progress set is calculated on the performance on the Twitter 2014 subset. For completeness, we report also the F measure calculated considering all the three classes in our model, including the neutral category 4.

The results are very encouraging: even if far from optimum, the system differs for only 3.29 points from the first ranked one (F=64.84). Furthermore, we observe that even if our system is trained only on tweets it is able to generalize on datasets from other domains, such as SMS and other microblogging services (i.e., LiveJournal). Conversely, the system performance drops on the Twitter 2014 Sarcasm set. This is consistent with results observed in our previous study (Basile and Novielli, 2014) on Italian tweets (Basile et al., 2014), where the 43% of misclassified negative cases were mostly ironic and would require common sense reasoning to detect the negative opinion expressed. Moreover a drop in perfomance on the sarcasm test set had been already

System	Positive			Negative			Neutral			F	
	P	R	F	P	R	F	P	R	F		
All features	85.42	55.30	67.13	60.51	52.05	55.96	61.11	86.93	71.77	64.95	
w/o keyword	88.23	49.81	63.67	59.62	51.78	55.43	59.18	89.16	71.14	63.41	(-2.37%)
w/o semantic	84.12	54.62	66.24	58.16	53.70	55.84	61,.28	85.61	71.43	64.50	(-0.69%)
w/o lexicons	83.31	52.89	64.70	60.92	39.73	48.09	58.14	58.14	70.00	60.93	(-6.19%)

Table 4: System results for all feature settings and all classes on the official test set Twitter 2015.

Test set	AVG (Fpos,Fneg)	Rank
Official 2015	61.55	12/40
Twitter 2014	65.11	25/40
LiveJournal 2014	70.05	-
SMS 2013	65.50	-
Twitter 2013	61.66	-
Twitter 2014 Sarcasm	37.30	-

Table 3: Task results.

observed for systems participating in the previous edition of the task (Rosenthal et al., 2014) and can be observed for all systems in the current edition. However, our system had a greater than average performance drop and we are currently studying this issue.

Observing the detailed scores for each class (first row of Table 4) we discover that the system performs better in the recognition of positive and neutral cases, in contrast with previous evidence from the experiment on the Italian corpus.

To further investigate the predictive power of the features in our model, we perform an ablation test on the Twitter 2015 test set, for which organizers provided the gold standard. We remove each group of features to assess the decrease of F measure on test data with respect to the setting including all features. Results are reported in Table 4 and demonstrate the importance of all feature groups.

Removing the sentiment lexicon group of features causes the highest decrease in performance. This is in contrast with previous evidence of our experiment on the Italian dataset of tweets, where a drop of performance of only 1% was observed. We provide a possible explanation to this by observing that only one sentiment lexicon was adopted in the study on the Italian dataset. On the contrary, in the current experiment on English tweets we can rely on a richer set of features due to the avaliablity of numerous lexicons, as explained in Section 2. Moreover, the Sentiment140 Lexicon and the Hashtag Sentiment Lexicon are both developed specifically to address sentiment analysis of tweets, thus providing higher coverage of lexical cues that are typical of microblogging.

Keyword and microblogging features are the second most useful group. This is consistent with evidence from the Italian experiment, for which we observe a comparable drop in performance on the polarity detection task. However, in the current experiment we also consider n-grams, which are not included in the feature set of the system for Italian. This consideration suggest that n-grams might contribute differently to the performance of sentiment classifiers depending on the language being used, thus suggesting directions for further investigation.

Finally, semantic features lead to the smaller drop in F measure when removed (-0.69%). This is in contrast with our previous findings in the Italian setting, where the semantic features plays a key role. This might be due to the prevalence of political topics in the Italian dataset, possibly causing a bias in our classifier due to the domain-specific lexicon about politics. This discrepancy indicates further directions for future investigation on the ability of semantic features in disambiguating polarity in microblogging, with respect to the topic being discussed and the language being used.

Future replications of this study will involve further data to validate and generalize our findings.

References

Pierpaolo Basile and Nicole Novielli. 2014. UNIBA at EVALITA 2014-SENTIPOLC Task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In *Proceedings of EVALITA 2014*, pages 58–63.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proc. of EVALITA 2014*, Pisa, Italy.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

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*, pages 168–177.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. 2014. TeamX: A Sentiment Analyzer with Enhanced Lexicon Mapping and Weighting Scheme for Unbalanced Data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632, Dublin, Ireland, August.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA, June.

Nicole Novielli and Carlo Strapparava. 2013. The Role of Affect Analysis in Dialogue Act Identification. *IEEE Transactions on Affective Computing*, 4:439–451.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, January.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86, July.

J. Pennebaker and M. Francis. 2001. Linguistic Inquiry and Word Count: LIWC. Erlbaum Publishers.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015, Denver, Colorado, June.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2):159–216, November.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in Short Strength Detection Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December.

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for Sentiment Analysis in Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2345–2354, Dublin, Ireland, 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.