UPF-taln: SemEval 2015 Tasks 10 and 11 Sentiment Analysis of Literal and Figurative Language in Twitter *

Francesco Barbieri, Francesco Ronzano, Horacio Saggion
Universitat Pompeu Fabra, Barcelona, Spain
name.surname@upf.edu

Abstract

In this paper, we describe the approach used by the UPF-taln team for tasks 10 and 11 of SemEval 2015 that respectively focused on "Sentiment Analysis in Twitter" and "Sentiment Analysis of Figurative Language in Twitter". Our approach achieved satisfactory results in the figurative language analysis task, obtaining the second best result. In task 10, our approach obtained acceptable performances. We experimented with both word-based features and domain-independent intrinsic word features. We exploited two machine learning methods: the supervised algorithm Support Vector Machines for task 10, and Random-Sub-Space with M5P as base algorithm for task 11.

1 Motivation

During the last decade the study and characterisation of sentiments and emotions in on-line user-generated content has attracted more and more interest. Since 2013 several tasks dealing with Sentiment Analysis have been organised in the context of SemEval. These tasks have been mainly focused on the analysis of short texts like SMS or tweets. In this paper we describe the approach adopted by UPF-taln team for tasks 10 and 11 of SemEval 2015, both dealing with the analysis of English tweets. Task 10 concerned "Sentiment Analysis in Twitter" and included different subtasks. We participated in the subtask B, named "Sentiment Polarity Classification". Given a message, we were asked to classify whether the message was of positive, negative, or neutral sentiment. In Task 11 the participants were asked to determine the polarity score (between -5 to +5) of tweets rich in metaphor and irony. Our model reaches satisfactory results in the figurative language task 11, however it has suboptimal performance in task 10.

We exploited an extended version of the tweet classification features and approach described in (Barbieri and Saggion, 2014). In particular, we experimented the use of intrinsic word features, characterising each word in a tweet to try to model and thus automatically determine its polarity. Thanks to intrinsic word features, we aimed to detect two aspects of tweets: the style used (e.g. register used, frequent or rare words, positive or negative words, etc.) and the unexpectedness in the use of words, particularly important for figurative language. We also exploited textual features (like word occurrences, bigrams, skipgrams or other word patterns) in order to capture the way words are used in positive and negative tweets. As machine learning approach we choose the supervised method Support Vector Machines (Platt, 1999) for task 10 and the regression algorithm Random-Sub-Space (Ho, 1998) with M5P (Quinlan, 2014) as base algorithm for task 11.

In Section 2 and 3 we describe the dataset used and the tools we employed to process the tweets. In Section 4 we introduce the features we built our model on. In Section 5 we discuss the performance of our model in SemEval 2015 and in Section 6 we

*The research described in this paper is partially funded by the Spanish fellowship RYC-2009-04291, the SKATER-TALN_UPF project (TIN2012-38584-C06-03), and the EU project Dr. Inventor (n. 611383).

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 704–708,
Denver, Colorado, June 4-5, 2015.

conclude with a recap of our approach and suggestions for further research.

2 Dataset

In order to train our systems we used in each task only the dataset provided by the organisers. For task 10 we were able to retrieve 9689 tweets, tagged as positive, negative and neutral (Rosenthal et al., 2015). For task 11 the dataset was a collection of 8000 figurative tweets annotated with sentiment scores from -5 to +5 (Li et al., 2015).

3 Text Analysis and Tools

In order to deal with the noisy text of Twitter we made use of the GATE application TwitIE (Bontcheva et al., 2013) where we modified the normaliser, adding new abbreviations, new slang words, removing URLs and changing the normalisation rules. Besides the tweet normalisation we also employed TwitIE for tokenisation, Part of Speech tagging and lemmatisation. We also used WordNet (Miller, 1995) to extract synonyms and synsets. We employed two sentiment lexicons, SentiWordNet3.0 (Baccianella et al., 2010) and the NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013) and two emotion lexicons NRC Hashtag Emotion Lexicon (Mohammad, 2012) and Depeche Mood (Staiano and Guerini, 2014). As frequency data for determining how often a word is used in English, we relied on the American National Corpus (Ide and Suderman, 2004); we also exploited the VU Amsterdam Metaphors Corpus (Steen et al., 2010) to find out how often a word is used in metaphors. Finally, the machine learning tool we used was Weka (Hall et al., 2009).

4 Our Method

We employed different machine learning methods for the two tasks. In task 10, as the classes were only three (positive, negative and neutral) we opted for a supervised learning method, and from our experiments with several classifiers, Support Vector Machines resulted to be the best one. On the other hand, in task 11 tweets were classified as belonging to one of 11 polarity classes associated with values ranging from -5 to 5, hence a regression approach was more suitable. The regression method employed was Random-Sub-Space with M5P as base algorithm. We also tried different mixed techniques, like using a supervised method to classify positive (0 to 5) and negative (-5 to 0), then a regression method (over the two subsets) but with no luck: pure regression methods fitted better task 11.

In both tasks we characterised each tweet using nine groups of related features all describing both intrinsic aspects of the words and word patterns. These groups of features are the following:

- Sentiments and Emotional Lexicons
- Frequency
- Lemma-Based
- Ambiguity
- Synonyms
- Adjective / Adverb Intensity
- Characters
- Part of Speech
- Bad Words

4.1 Sentiments and Emotional Lexicons

Using sentiment lexicons in Sentiment Analysis has been a common and rewarding practice (Mohammad et al., 2013; Kiritchenko et al., 2014). The characterisation of the sentiment associated to words in tweets is important for two reasons: to detect the *global sentiment* (e.g. if tweets contain mainly positive or negative terms) and, in the case of figurative language, to capture *unexpectedness* created by a negative word in a positive context and viceversa. Using the two sentiment lexicons and two emotional lexicons mentioned in Section 3, we computed the *number of positive / negative words*, the *sum of the intensities of the positive / negative scores of words*, the *mean of positive / negative score of words*, the *greatest positive / negative score*, the *gap between the greatest positive / negative score and the positive / negative mean*. These features are computed including all the words of each tweet. We also determined these features by considering separately Nouns, Verbs, Adjectives, and Adverbs (we calculate the features by considering only words characterised by a specific Part of Speech).

4.2 Frequency

To design the Frequency feature we used two frequency corpora: the American National Corpus and the VU Amsterdam Metaphors Corpus. From these corpora we extracted three features: *rarest word frequency* (frequency of the rarest word included in the tweet), *frequency mean* (word frequency arithmetic average) and *frequency gap* (the difference between the two previous features). As previously done, we computed these features by considering only Nouns, Verbs, Adjectives, and Adverbs.

4.3 Lemma-Based

We designed this group of features to detect common word-patterns in positive and negative tweets. The lemma-based features are three: *lemma+pos* (the combination of each lemma and its Part of Speech in the tweet), *bigrams* (combination of two lemmas in a sequence) and *skip one gram*, combination of two lemmas with distance one (two lemmas separated by one lemma).

4.4 Ambiguity

Ambiguity is modelled with WordNet. Our hypothesis is that if a word has many meanings (synset associated) it is more likely to be used in an ambiguous way. For each tweet we calculated the *maximum number of synsets* associated to a single word, the *mean synset number* of all the words, and the *synset gap*—the difference between the two previous features. We determine the value of these features by including all the words of a tweet as well as by considering only Nouns, Verbs, Adjectives or Adverbs.

4.5 Synonyms

We carried out an analysis of the choice of synonyms as follows: for each word in the tweet we retrieve its list of synonyms, then we computed, across all the words of the tweet: the *greatest / lowest number of synonyms* with frequency higher than the one present in the tweet, the *mean number of synonyms* with frequency greater / lower than the frequency of the related word present in the tweet. We determine also the greatest / lowest number of synonyms and the mean number of synonyms of the words with frequency greater / lower than the one present in the the tweet (*gap* feature). We computed the set of Synonyms features by considering both all the words and also restricting the calculation to words with the Part of Speech tags as above.

4.6 Adjective / Adverb Intensity

Using the Potts (2011) intensity scores of Adjectives and Adverbs, we calculated three features: the *most intense* adjective/adverb and the *intensity mean* of the adjective/adverb of the tweet.

4.7 Characters

We also wanted to capture the punctuation style of the author of a tweet. Punctuation and type of characters used are very important in social networks: a full stop at the end of a subjective message may change the polarity of the message. Each feature is a count of specific punctuation marks, including: ".", "#", "!", "?", "$", "%", "&", "+", "-", "=", "/". Moreover we count as well number of *uppercase* and *lowercase* character.

4.8 Part of Speech

The features included in the Part of Speech group are designed to capture the structure of positive and negative tweets. The features of this group are eight and each one of them counts the number of occurrences of words characterised by a certain Part of Speech. The eight Part of Speech considered are *Verbs*, *Nouns*, *Adjectives*, *Adverbs*, *Interjections*, *Determiners*, *Pronouns*, and *Appositions*.

4.9 Bad Words

Since Twitter messages often include *bad words*[1], we count them as they may be used more often in negative messages.

5 Experiments and Results

In this section we present our results in the two tasks (see Table 1 and Table 2). We only report final results (mean of Precision, Recall and F-Measure of each class), for more details please refer to the task 10 and task 11 papers (Rosenthal et al., 2015; Li et al., 2015).

[1]We enriched with more variants this list: https://github.com/shutterstock/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

5.1 Task 10-B

Given a message, classify whether the message is of positive, negative, or neutral sentiment. Our model scores at position 27th out of 40 groups. Systems were evaluated with the mean of the F-measures of Positive, Negative and Neutral classes. Our score is 9 points less than the best system. A considerable number of tweets in the test set were considered sarcastic tweets complicating the sentiment analysis task. With this test subset our system improves its performances globally scoring at the 11th position. See Table 1 for the results in each test set. The features that perform better are from the group Sentiments and Emotion Lexicons, that achieve information gain scores of 0.133. Even if less influent, the Frequency group obtains a score of 0.09. The other group of features are not very important for this task, and the information gain scores are less than 0.3.

	F-Measure	Rank
Twitter 2014	65.05	27th
Sarcasm	50.93	11th
Twitter 2013	66.15	17th
SMS 2013	57.84	31st
LiveJournal 2014	64.5	31st

Table 1: Task 10 results. For each test set we report F-Measure and ranking comparing to other systems.

5.2 Task 11

Given a set of tweets that are rich in metaphor and irony, the goal is to determine whether the user has expressed a positive, negative or neutral sentiment in each, and the degree to which this sentiment has been communicated.

A vector space model was used to evaluate the similarity of the predictions of each participating system to the human-annotated gold standard. The list of expected gold-standard sentiment scores was used to construct a normalised gold-standard vector, while a comparable vector will be constructed from the predictions of a participating system. The cosine distance between vectors was then used as a measure of how well the participating system estimates the gold-standard sentiment scores for the whole of the test set (Li et al., 2015).

In this task our model ranked second out of 15 participants. We obtained a cosine similarity of 0.710 and a Mean Squared Error (MSE) of 2.458. The best system cosine and MSE scores were respectively 0.758 and 2.117. In Table 2 the reader can find all the results.

In Table 3 we show experiments to analyse the contribution of each type of feature to the final results. The most important contribution is given by the Sentiment lexicons NRC and SentiWordNet (see Section 4.1). Also the Synonyms feature is important with a cosine similarity of 0.564. The feature that was less influent to the final classification was Intensity of Adjectives and Adverbs.

	MSE	Cosine
Overall	2.458	0.711
Sarcasm	0.934	0.903
Irony	1.041	0.873
Metaphor	4.186	0.520
Other	3.772	0.486

Table 2: Task 11 results measured by the Cosine Similarity and the Mean Square Error over the test set (Overall) and for its subsets: sarcasm, irony, metaphor and other (non-figurative tweets).

Feature	Cosine Similarity
NRC H. Sentiment	0.578
SentiWordNet	0.562
Synonyms	0.564
Characters	0.550
Part of Speech	0.550
Depeche Mood	0.550
Lemma-Based	0.547
NRC H. Emotion	0.547
Bad Words	0.547
Frequency	0.546
Ambiguity	0.546
Intensity	0.544

Table 3: Task 11 contribution of each group of feature. The best feature group was Sentiment, in particular the features computed with the NRC Hashtag Sentiment Lexicon, see Section 4.1.

6 Conclusions

In this paper we have described our participation to the SemEval task 10 and 11. Besides the word-

based features, we experimented the use of intrinsic word features to characterise positive and negative tweets. In task 10 our system obtains average performances leaving room for important improvements to our approach. Our system obtains very good results in task 11, ranking second out of 15 participating teams. The difference in performance in the two tasks was expected since our model is the adaption to sentiment analysis of a model for irony (Barbieri and Saggion, 2014) and sarcasm (Barbieri et al., 2014) detection in Twitter, thus it fits better the figurative language identification task. Yet, both models can be improved and we are planning to add new features (vector space models and distributional semantics among others) and experiment new machine learning techniques (e.g. cascade classifiers for task 10 or different regression algorithms for task 11).

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Francesco Barbieri and Horacio Saggion. 2014. Modelling Irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April.

Francesco Barbieri, Horacio Saggion, and Ronzano Francesco. 2014. Modelling sarcasm in twitter, a novel approach. *ACL Workshop on Sentiment Analysis: WASSA*.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. 2013. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of Recent Advances in Natural Language Processing Conferemce*.

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

Tin Kam Ho. 1998. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844.

Nancy Ide and Keith Suderman. 2004. The American National Corpus First Release. In *Proceedings of the Language Resources and Evaluation Conference*.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, August.

Guofu Li, Aniruddha Ghosh, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. 2015. Task 11: Sentiment Analysis of Figurative Language in Twitter. Denver, Colorado, USA, June, 4-5.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

Saif Mohammad. 2012. #Emotional Tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, Canada, 7-8 June.

John Platt. 1999. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. *Advances in kernel methodssupport vector learning*, 3.

Christopher Potts. 2011. Developing adjective scales from user-supplied textual metadata. *NSF Workshop on Restructuring Adjectives in WordNet. Arlington,VA*.

J Ross Quinlan. 2014. *C4. 5: programs for machine learning*. Elsevier.

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, USA, June.

Jacopo Staiano and Marco Guerini. 2014. DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. In *52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, page 427433, Baltimore, Maryland, USA,, June.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.