POPSTAR at RepLab 2013: Polarity for Reputation Classification

João Filgueiras and Silvio Amir

INESC-ID, Technical University of Lisbon
{jfilgueiras,samir}@inesc-id.pt

Abstract. This paper describes our participation in the Polarity for Reputation classification task of RepLab 2013. Our system leveraged on a set of components previously developed for a Twitter message polarity classifier. Following a supervised approach, a Logistic Regression classifier is trained from annotated data. A refined language model is used to represent tweets in terms of a vocabulary consisting only of the most informative terms with word features weighted using a measure from the Information Retrieval field. To help reduce the sparseness of the feature vector, the model is enriched with another, more compact, representation of the words. Finally, we extract features to capture the use of informal and affective language. Our approach ranked in the top three for all the metrics, showing that the strategies for Twitter Sentiment Analysis are useful for the task of Polarity for Reputation classification.

1 Introduction

Twitter has become a vast repository of user-generated content over the years. A current research trend is to use this repository, and others like it, as a source of indicators of public opinion on a number of topics. Some popular examples are the prediction of elections or the prediction of box-office revenues for movies.

The RepLab Polarity for Reputation Classification task aims to use tweets as an indicator of the reputation of entities, such as companies, brands or artists. To accomplish this goal the reputation analysis problem is divided into four tasks: filtering, polarity, clustering and priority. The filtering step removes tweets that are not related to an entity. The polarity step rates the impact of the tweet on the reputation of the entity as positive, negative or neutral. The clustering step, aims to cluster tweets pertaining to an entity into topics and finally, the priority step assigns a level of importance to each topic. RepLab participants were given an annotated dataset and asked to implement one or more components of this system.

It is clear that reputation analysis and sentiment analysis are different tasks. Objective facts can still have a negative impact on the reputation of an entity, and tweets expressing negative sentiment may still be positive for a reputation or vice-versa. However, from a technical perspective, how different are the underlying problems? And to what extent are the techniques suitable for one task effective on the other? Motivated by these questions we participated in RepLab 2013 using an effective classifier we had developed for a Twitter message polarity classifier. This classifier was tailored to participate in the task proposed in the Sentiment Analysis in Twitter track of SemEval 2013,

a workshop focused on the evaluation of semantic analysis systems (Nakov et al., 2013) and achieved state-of-the-art results.

The remainder of the document is organized as follow: next, the annotated dataset is presented, in Section 2 we describe our approach in detail, Section 3 shows the results of the experiments and we briefly conclude in Section 4.

1.1 Dataset

The RepLab 2013 organization provided an annotated dataset that contained a total of 33,191 tweets annotated for polarity. Of these, 7,073 were written in Spanish and 26,118 in English. The content of webpages linked by URLs in each tweet was also provided. The tweet distribution across classes is shown in Table 1.

Language	**Positive**		**Negative**		**Neutral**	
English	15545	59.5%	3429	13.1%	7144	27.4%
Spanish	3195	45.2%	2143	30.3%	1735	24.5%

Table 1. Tweet distribution across polarity classes.

2 Approach

2.1 Overview

Following a supervised approach, we used the annotated data to train a Logistic Regression classifier. Each tweet was modelled as a feature vector consisting of a bag-of-words vocabulary representation. However, the lexical variation introduced by typos, abbreviations, slang and unconventional spelling found in Twitter data, leads to very large vocabularies. The resulting sparse vector representations with few non-zero values hamper the learning process. To overcome this problem, words with high entropy (not discriminative of any of the classes) were discarded and Brown Clusters (Brown et al., 1992) were used as a complementary representation of the tweet to enrich the feature vector.

We used sentiment lexicons to extract features based on the prior polarity of words, taking the presence of negation particles into consideration. Some microblog oriented features were included to capture particular aspects of this type of text (e.g. presence of emoticons). Finally, the title of the web pages referred in the messages was included to provide additional context.

To deal with the fact that there were two different languages, we chose to train a separate classifier for each language. The Spanish classifier was a simplified version of the English classifier, without Brown clusters and lexicons features.

Each major component of our system is described in detail in the following sections.

2.2 Preprocess

In order to cope with the noisy content and reduce the vocabulary size, the following preprocess steps were taken: stop words were discarded, user mentions (*@username*) were replaced with a fixed tag <USER> and URLs with the tag <URL>. Then, messages were normalized by converting to lower-case and reducing character repetitions to at most 3 characters (e.g. *"hellooooooo!"* would be normalized to *"hellooo!"*). Finally, words were stemmed using the Snowball Stemmer implementation of NLTK [1].

2.3 Document Representation

Negation The presence of a negation word can have great impact in the meaning of a sentence, e.g., the expressions *"very good"* and *"not very good"* convey opposite sentiments. Therefore, following the work of Pang et al. (2002), negation was directly integrated in the words representation. All the words between a negation word and the first punctuation mark, were suffixed with the _**NEG** tag. The list of negation words was compiled manually.

Weighting Schemes for Word Features Typical schemes proposed for weighting word features in text classification tasks are binary weighting, term frequency and tf.idf. However, Paltoglou and Thelwall (2010) showed that advanced weighting schemes used in Information Retrieval can enhance sentiment classification accuracy. As these measures capture the relative importance of a term in two classes (positives and negatives in this case), they provide more informative word weights for the task at hand. We found experimentally that delta-tf.idf weight function yields the best results.

Brown Clusters The Brown algorithm is a hierarchical clustering algorithm that clusters words to maximize the mutual information of bi-grams. The hierarchical nature of the clustering allows words to be represented at several levels in the hierarchy, which can compensate for poor clusters of a small number of words. Brown clusters are created by applying the Brown algorithm to a large corpora, capturing relations between bi-grams to form a denser representation of a vocabulary. Using these clusters we also represented documents in terms of a more compact vocabulary, where each word was mapped to its corresponding cluster. Plugging these clusters as extra-features into the document model, can alleviate the problems of feature vector sparseness and unseen words.

Word Entropy Previous work has shown that in order to improve classification performance when using bag-of-words, words with high entropy, i.e., that do not contribute strongly to any of the classes should be discarded (Pak and Paroubek, 2010). The entropy of the probability distribution of a word appearing in the different classes was computed using the Shannon and Weaver (1948) definition.

[1] `http://nltk.org/`

A high entropy value indicates that a word appears evenly in all the classes, whereas low entropy values mean that a word is more frequent in one of the classes, hence, being more discriminative of a given sentiment. After computing the entropy values for each term, a threshold τ was defined and words with entropy above τ were discarded, reducing vocabulary size.

2.4 Features

The document representation model was enriched with features that take into account the presence of words with prior polarity, such as *"happy"* (positive) or *"sad"* (negative) in a document. In the spirit of previous approaches, features that aim at capturing the creative and informal use of language in *tweets*, were also extracted. In summary, the following features were employed:

- **Sentiment Lexicons:** number of words with positive\negative prior polarity and a score obtained by summing both. Negation was taken into account by detecting the presence of a negation token in a window of two words. Bing Liu's Opinion Lexicon[2] was employed for this feature.
- **Heavy Punctuation:** number of sequences of exclamation marks, question marks and combinations of both.
- **Upper-case Words:** number of words all in upper case.
- **Emoticons:** number of positive and negative emoticons. The polarity of emoticons was assessed with custom regular expressions and a list of polar emoticons used in SentiStrength.
- **Emphasized Words:** number of words emphasized with more than 2 character repetitions, e.g., *"awesoooome"*.

2.5 Context

As mentioned in Section 1.1 the training and test datasets also contained the contents of web pages linked on tweets. To take advantage of this extra information we employed a simplistic strategy. We extracted the title HTML tag of these pages and appended it to text of the tweet associated with it.

3 Results

To test the impact of each feature and determine the best settings for this problem, we used a 70%-30% split to evaluate different versions of the classifier. Table 2 reports our results for the English classifier. The baseline used only a binary bag-of-words and Brown clusters representation of each tweet without any vocabulary reduction. We then added the traditional features mentioned in Section 2.4. The binary weighting scheme was changed afterwards for a Delta-tf.idf, followed by a vocabulary reduction using entropy, and finally the addition of the HTML title tags.

[2] `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

Classifier	Accuracy	Polar F1
Baseline	71.7%	69.5%
+ Features	72.4%	71.0%
+ Delta TF.IDF	72.8%	71.0%
+ Vocabulary Reduction	73.5%	71.5%
+ HTML Title	73.7%	72.0%

Table 2. Experimental results for the English classifier iterations.

Table 3 shows a simplified table for the Spanish classifier. In this case the baseline was simply a binary bag-of-words. As mentioned previously, the full classifier employed every strategy in the English classifier except Brown clusters and lexicons.

Classifier	Accuracy	Polar F1
Baseline	71.6%	74.5%
Full classifier	72.3%	75.5%

Table 3. Experimental results for the Spanish classifier.

Finally, Table 4 shows our results in the final evaluation by the organization of RepLab 2013, using a test dataset. Each participant could submit up to ten different runs, and we opted to submit different variations of the vocabulary reduction and the use of the context HTML title tags. Several vocabulary sizes were tested experimentally and the best were submitted. The reported metrics of reliability and sensitivity, related to precision and recall of the positive and negative classes, are described by Amigo et al. (2012).

Run	English Vocab.	Spanish Vocab.	Title	Accuracy	Pearson Corr.	Reliability	Sensitivity
1	80%	30%	No	63.2%	0.883	42.6%	33.6%
2	98%	80%	No	63.3%	0.881	42.0%	32.8%
3	80%	80%	No	63.6%	0.883	43.0%	33.4%
4	80%	30%	Yes	63.6%	**0.889**	43.0%	**34.0%**
5	80%	80%	Yes	**63.9%**	0.888	**43.3%**	33.9%

Table 4. Submitted runs.

Our submissions ranked third in the accuracy metric, first in Pearson correlation and second in the F-measure of sensitivity and reliability.

4 Conclusions

Although the problems of evaluating the polarity of sentiment and analyzing polarity for the reputation of an entity expressed in a tweet are different we found that the same principles and techniques can be used.

Our classifier achieved good results even in a different language, such as Spanish, without any language-specific features. The experiments and evaluation show that the proposed approach is robust and indicate that the underlying problems of reputation analysis are not very different from free domain sentiment analysis.

We also note that the use of context, even in such a simplistic way as the one we used, improves the overall results. It becomes clear that more sophisticated approaches should be explored in the future.

Acknowledgments

This work was partially supported by FCT (Portuguese research funding agency) under project grants UTA-Est/MAI/0006/2009 (REACTION) and PTDC/CPJ-CPO/116888/2010 (POPSTAR). FCT also supported scholarship SFRH/BD/89020/2012. This research was also funded by FCT under contract Pest-OE/EEI/LA0021/2013.

Bibliography

Amigo, E., Gonzalo, J., and Verdejo, F. (2012). Reliability and sensitivity: Generic evaluation measures for document organization tasks. Technical report, Technical report, UNED.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*. ACM.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC*, pages 1320–1326.

Paltoglou, G. and Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Shannon, C. E. and Weaver, W. (1948). A mathematical theory of communication.