Detecting Entity-Related Events and Sentiments from Tweets Using Multilingual Resources

Alexandra Balahur and Hristo Tanev
alexandra.balahur@jrc.ec.europa.eu
hristo.tanev@ext.jrc.ec.europa.eu

European Commission Joint Research Centre
IPSC, GlobeSec, OPTIMA
Via E. Fermi 2749, Ispra, Italy

Abstract. This article presents the details of the participation of the OPTAH team to the CLEF 2012 RepLab profiling (polarity classification) and monitoring tasks. Specifically, we present the manner in which the OPAL system has been modified to deal with opinions in tweets and how the use of rules involving the language use in social-media can help to achieve good results as far as polarity classification is concerned, even in a language for which we have just a small polarity lexicon. Additionally, we show how we can employ the values computed for sentiment intensity (especially the negative ones) to classify the importance of event-related clusters of tweets. Our methods, although quite simple, obtained promising results in the RepLab evaluations.

1 Introduction

In the new Social Web era, the influence the user-generated contents have on all spheres of life has reached an unprecedented level. People's comments on news, events and their personal opinions on persons and companies worldwide have made the Internet a rich source of information, highly relevant for the people or companies in question and their stakeholders. Online Reputation Management deals with the issue of detecting and employing "positive" and "negative" clues expressed in online contents on such people and companies, in an automatic manner. As stated by Balahur [4], this task is highly complex, as it deals with important issues in opinion mining, sentiment analysis, bias detection, Named Entity discrimination, online trust and reputation management, topic modeling, good versus bad news classification and other aspects which, in themselves are not trivial in Natural Language Processing. This article presents the details of the participation of the OPTAH team to the CLEF 2012 RepLab profiling (polarity classification) and monitoring tasks. The main objectives of our experiments were:

1. For the polarity task:
   - test if the methods we have developed for sentiment analysis for other text types can be adapted to the case of tweets (short texts) and what changes are required to that aim;
   - to test and evaluate, in comparison, a semi-supervised versus an unsupervised method for sentiment analysis in this type of texts; and

   - to measure the impact of resources that are typical of social media - e.g. collections of smileys, colloquial expressions, slang, repetitions of punctuation signs, etc. and the use of an algorithm to normalize words can help to more accurately detect opinions in tweets.
2. For the monitoring task:
   - to test how well a clustering method that has initially been employed in the case of news can be customized to deal with news reported in tweets; and
   - to test in how far we can employ the intensity of the sentiment detected in the tweets within clusters to sort them depending on their priority.

In the first case, although the adaptation to the Twitter domain was not very extensive in terms of sentiment-bearing words, our results showed that the use of rules taking into account the typical phenomena in the short informal texts can achieve good results, our two submissions ranking 8th and 9th overall and achieving the 5th rank in the case of Spanish, in terms of F-score for polarity and the second rank in terms of polarity accuracy for Spanish. In the second task, we could see that the negativity expressed in the comments were important to the priority of the clusters that contained those comments. Nevertheless, additional methods to score the "negativity" of news have to be employed, as well as the added use of "good" versus "bad" news terms, which were disregarded by the OPAL system.

2 Profiling Task - Polarity Classification

For the polarity classification task, we employed to approaches. The first one was semi-supervised, using a variant of the OPAL system [5], whose extension is presented in the following section. The second one was unsupervised, using only lexicons of words that relate to polarity, as well as a set of rules for modifiers and negation. The two approaches are described in the following subsections. In order to prepare the tweets for analysis, the texts were tokenized and subsequently the tokens were preprocessed as follows:

1. Word normalization. The words in the tweets were compared against the words in the Roget's Thesaurus. Subsequently, words that were not found in the dictionary were processed, eliminating repeated letters until they matched a word in the dictionary. The words were also matched against the affect lexicons we employed in our method, which were The General Inquirer [8] list of sentiment words, the Linguistic Inquiry and Word Count - LIWC - [9] resource and MicroWNOp [7] as well as the dictionary obtained by Steinberger et al. [6] for Spanish. This is important, as for the second method, which is based on the polarity and intensity values of concepts, the value of the word that is "stressed" by writing it with repeated letters receives an increment in polarity (i.e. for positive words, 1 is added to the total polarity value and for negative words, 1 is subtracted from the total polarity value).
2. Emoticon replacement. We employed an emoticon dictionary and replaced the emoticons found in the tweets with the word they signify (e.g. ":)" is replaced with "happy").

3. Repeated punctuation sign normalization. In the tweets, we reduced multiple punctuation signs to only one and, for the second approach, added or subtracted 1 from the total polarity value.

2.1 OPAL - a System for Opinion Detection from Text

This run was submitted with the acronym OPTAH_1.

In order to determine the polarity of the sentences, we passed each sentence through an opinion mining system employing SVM machine learning over the NTCIR 8 MOAT corpus - for English and the Spanish translation, obtained by Balahur and Turchi [11] -, the MPQA corpus for English, EmotiBlog [10] for English and Spanish and the tweets given in the training set by the organizers of the RepLab 2012 competition. As opposed to the system employed in the NTCIR MOAT 8 task [5], we only used tokenization and did not perform any parsing, as tweets are many of the times not fully-formed sentences. Each of the positive, negative, negation and modifier (intensifier, diminisher) words found in this corpora were matched against the General Inquirer, Opinion Finder, MicroWordNet and LIWC resources anreplaced by the "POSITIVE", "NEGATIVE", "NEGATOR", "INTENSIFIER" and "DIMINISHER" labels. Subsequently, we represented the sentences in the training set as a vector containing the presence (1) or absence (0) of all the unigrams and bigrams in the corpora used for training. With the vectors thus obtained, we employed the Support Vector Machines implementation in Weka (the SMO version) and created a learning model. The tweets in the test set were represented as vectors whose features corresponded to the presence or absence of the unigrams and bigrams in the training sets.

2.2 Opinion Detection from Text Using Opinion Lexica and Rules

This second run was submitted with the acronym OPTAH_2.

In this second approach, we employed a simpler method to compute the polarity and intensity scores. Each of the sentiment lexicons employed were mapped to 4 values of polarity - high positive (with a value of 4) , high negative (-4), positive (1), negative (-1). Additionally, we added slang words for both languages (e.g. "LOL" with a value of 4, "joder" with a value of -4). Additionally, we employed a set of rules, to take into consideration negation, modifiers, repeated punctuation signs and emoticons, as follows:

- Negation treatment. When a negation was found, the polarity of the subsequent sentiment-bearing words found in the tweet was inverted. We excluded the known cases of "false negations", such as "not only", "no solamente".
- Modifier treatment. When an intensifier was found, the polarity of the follow sentiment-bearing word in the tweets was multiplied with 1.5. In the case of diminishers, the polarity of the sentiment bearing word that followd it, the value of its polarity was multiplied with 0.5.
- Emoticon treatment. When an emoticon is found, the score it is given is of the word that it represents (e.g. ":(" has the value -1, of "sad").

- Repeated letters treatment. When a word has repeated letters and it is found in the polarity lexicon, its polarity value is multiplied by 1.5.
- Repeated punctuation signs. In the case of exclamation signs, the value of the entire sentence preceding it is multiplied by 1.5. In the case of fullstops, the value of the preceding sentence is multiplied by 0.5.

2.3 Results and Discussion

For the two runs we submitted, we obtained the following results, in terms of polarity accuracy, R polarity, S polarity and F-score of R and S polarity, respectively: OPTAH_1 (0.3644, 0.3256, 0.3102, 0.3048), OPTAH_2 (0.3705, 0.4048, 0.2689, 0.3042), scoring 8th and 9th of 34 runs in terms of F(R,S). Per language, for English, the results, in the same order, were: OPTAH_1 (0.3207, 0.3050, 0.2920, 0.2810) and OPTAH_2(0.3293, 0.4061, 0.2523,0,2922).

For Spanish, the results were: OPTAH_1 (0.4430, 0.3041, 0.2901, 0.2837) and OPTAH_2 (0.4435, 0.3695, 0.2567, 0.2844). We can see that for the case of English, using more resources deteriorated the performance and the use of the semi-supervised method of learning actually produced worse results than the use of a simple, lexicon and rule-based system. In case of Spanish, our systems ranked among the first three in terms of accuracy and F-measure, showing that a smaller, but more precise lexicon (containing also slang), combined with a set of rules that capture the manner in which expressions of sentiment are stressed upon in Social Media, can better help to classify tweets.

3 Monitoring Task

We participated in the RepLab monitoring task. In this task we used multilingual lists of keywords, extracted my Europe Media Monitor [1], which were used as features for clustering. Then, we used the second system for sentiment detection described above (OPTAH_2) to define the priority of the clusters. Our assumption was that clusters, which convey negative opinions should be considered more relevant for reputation management, since they may report about major issues, related to the mentioned organization.

Our algorithm has two stages of processing: clustering and priority definition. Now we will explain in more details each of these steps.

Clustering is performed in three steps: First, for each tweet, we build a vector from the Europe Media Monitor keywords which appear in this tweet; doing this we ignore very frequent keywords. Each dimension of our vector corresponds to one word which appears in the tweet. The values of the vector components are defined, using log-likelihood ratio, considering probability of appearance of the word in a large news corpus of 100'000'000 words. The fact, that we used a news corpus and not one derived from tweets influences the accuracy of our approach; however, we did not have Twitter-specific keywords.

Then, we count which of the Spanish or English keywords are more represented in the tweet and consider the tweet as English or Spanish, according to the language, from

which the majority of the keywords come. We did not consider tweets with less than 3 keywords, since these were most probably not informative.

Finally, we cluster the tweet's vectors, using agglomerative clustering with a threshold, previously optimized on the training set of clusters. Our criteria for optimization was that the average reliability and sensitivity of the clustering for the training set entities are balanced. In our experiments we used the CluTo clustering tool [2].

We defined the priority of the clusters by using sentiment detection. We assumed that negative tweets convey information about issues and problems, related to the organization of interest and its products or services. Negative opinions are important for reputation management, since negative perception of an organization can be exploited against it by its competitors. Also, by analysing the negative opinions, the organization could find its weak points as seen by people and improve its image. For example, one of the negative tweets about Blackberry was:
So my blackberry broke again this morning, and is not working again
Similar opinions should be important for Blackberry, since they show problems with these products. Or let's consider the following example:
Bank of America bugs the shit out of me in general. 20 years! I think I am a financial masochist
Again, here we have a negative opinion about Bank of America and its services. Tweets, similar to this one should be considered important for Bank of America, since they show potential weaknesses in their services. Another example for the same organization:
Bank of America refusing to do business with certain companies... WOW for a bank that nearly went bankrupt and closing branches all over
This tweet directly states a negative fact about Bank of America, at the same time it shows clearly negative attitude towards the bank.

In order to detect negative tweets, we run our multilingual system OPTAH_2 and we detected the clusters, which contain negative tweets. These clusters were considered important and their priority was set to alert level , while the clusters not containing negative tweets we considered unimportant and their priority level was set to average.

One of the weaknesses of our clustering approach was that we used clustering based on purely lexical features. We could have considered for example, synonyms and similar words. Also, using dimentionality reduction, clustering can be done on a reduced feature space [3]. This could potentially result in a better clustering. Another possibility was to calculate a table of distributional similarity between the frequent keywords. In this way, we could overcome the restrictions of lexical similarity.

Another possibility to improve the results lays in improving calculation of cluster priority. Currently, we used only sentiment detection. One could use also the size of the cluster, it's lexical content, also the fact that some tweets are retweeted or replied-to, how many Web links are provided in the tweets, etc., in order to calculate better the priority. This can be formulated as a supervised machine learning task, where certain tweets or clusters are marked manually with their level of priority and the features are the previously mentioned characteristics.

We submitted one run for the monitoring task. It was ranked in the middle of the ranked list of runs, with the following scores: 0.7 for R CLUSTERING (BCubed precision), 0.34 for S CLUSTERING (BCubed recall), 0.38 for F(R,S) CLUSTERING,

0.19 for R PRIORITY, 0.16 for S PRIORITY, 0.16 F PRIORITY, 0.37 R, 0.19 S and an overall F(R, S) of 0.22. Considering the simplicity of our approach, we consider the results satisfactory, still possible to improve. One of the main problems was that negative sentiment alone was not enough to detect important tweet clusters.

4 Conclusions and Future Work

From our experiments with the training data and from the results obtained in the competition, we could see that indeed reputation management is a difficult task. The main challenge is related to the language used in social media, the shortness of texts, the assumed knowledge on the context (i.e. people use hashtags to refer to specific events, which are presented in traditional media) and the difficulty of assessing "good" and "bad" news from the perspective of different domains. As future work, we plan to use the EMM categories for events as additional clues to the positivity and negativity of events and develop a method to detect topic-specific types of events, which we can then classify in terms of positive or negative impact on the entity in dependence to the opinion expressed in social media. Further on, we will extend on the method developed by (Tanev et al., 2012) to link tweets to news and thus be able to explore a higher quantity of text for the reputation management task.

References

1. Steinberger, R., Pouliquen, B., Van der Goot, E. 2009. An Introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando, Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.
2. Karypis, G. CLUTO. Available online: http://glaros.dtc.umn.edu/gkhome/views/cluto/.
3. Song, W. and Park, S.C. 2007. A novel document clustering model based on latent semantic analysis. In Proceedings of the Third International Conference on Semantics, Knowledge and Grid, pages 539 -542.
4. Balahur, A.. 2012. The Challenge of Processing Opinions in Online Contents in the Social Web Era Workshop Language Engineering for Online Reputation Management at LREC 2012.
5. A. Balahur, E. Boldrini, A. Montoyo, P. Martinez-Barco. 2010. The OpAL System at NTCIR 8 MOAT. Proceedings of NTCIR 8, 2010.
6. Steinberger, J., Lenkova, P., Ebrahim, M., Ehrman, M., Hurriyetoglu, A., Kabadjov, M., Steinberger, R., Tanev, H., Zavarella, V., Vazquez, S.. 2011. Creating Sentiment Dictionaries via Triangulation. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011).
7. Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., Gandini, G.. 2007. Language resources and linguistic theory: Typology, second language acquisition, English linguistics, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.
8. Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M.. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press.
9. Pennebaker, J.W., Francis, M.E., Booth, R.J.. 2001. Linguistic Inquiry and Word Count: LIWC2001. Mahwah, NJ: Erlbaum Publishers.

10. Boldrini, E., Balahur, A., Martinez-Barco, P., Montoyo, A.. 2010. EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. Proceedings of the 4th Linguistic Annotation Workshop (LAW IV), 2010.
11. Balahur, A., Turchi, M.. 2012. Multilingual Sentiment Analysis Using Machine Translation?. Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2012).