DLSI-Volvam at RepLab 2013: Polarity Classification on Twitter Data

Alejandro Mosquera[1,2], Javi Fernández[1],
José M. Gómez[1], Patricio Martínez-Barco[1], and Paloma Moreda[1]

[1] Department of Software and Computing Systems,
University of Alicante, Alicante, Spain
http://www.dlsi.ua.es
[2] Volvam Analytics Ltd., Dublin, Ireland
http://www.volvam.com
{amosquera,javifm,jmgomez,patricio,paloma}@dlsi.ua.es

Abstract. This paper describes our participation in the profiling (polarity classification) task of the RepLab 2013 workshop. This task is focused on determining whether a given text from Twitter contains a positive or a negative statement related to the reputation of a given entity. We cover three different approaches, one unsupervised and two unsupervised. They combine machine learning and lexicon-based techniques with an emotional concept model. These approaches were properly adapted to English and Spanish depending on the resources available for each language. We obtained promising results in the overall evaluations, reaching a F-score of 34% and a sensitivity of 40% in the best cases. The reasonable level of performance compared to other methods encourages us to continue working on the improvement of the proposed approaches.

Keywords: online reputation, sentiment analysis, polarity classification, text normalisation, machine learning, lexicon, emotion concepts

1 Introduction

Nowadays, *social media* applications have allowed users to have an active participation through their comments and opinions, stated about a wide range of topics and services. This subjective information is very valuable because it determines the reputation of public figures and companies in the marketplace of personal and business relationships. However, it is not feasible to monitor this information in a manual way, because the amount of information is very large and is updated very quickly. Therefore, automatising this process is essential. The field of *on-line reputation management* (ORM) studies automated ways to track the opinion of the users about qualitative or quantitative aspects dealing with several challenges such as subjectivity, textual noise or domain heterogeneity. This task is very complex, as it deals with important issues in opinion mining, sentiment analysis, bias detection, named entity discrimination, topic modelling and other aspects which are not trivial in *natural language processing* [1].

RepLab 2013 is a competitive evaluation exercise for ORM systems, focusing in monitoring the reputation of entities (companies, organisations, celebrities, etc.) on *Twitter*[3] [2]. In this article we focus our participation in the profiling (polarity classification) task. The goal of this task is to decide if the tweet content has positive or negative implications for the reputation of a given entity. Polarity for reputation is substantially different from standard sentiment analysis, because the goal is to find what implications a piece of information regardless of whether the content is opinionated or not. In addition, negative sentiments do not always imply negative polarity for reputation and vice versa (e.g. `"R.I.P. Whitney Houston. We'll miss you"` has a negative associated sentiment but a positive implication for the reputation of Whitney Houston).

We propose three different approaches to face this task. Our first approach is unsupervised and makes use of *fuzzy lexicons* in order to catch informal variants that are common in Twitter texts. The second one is supervised and extends the first approach with *machine learning* (ML) techniques and an *emotion concept model*. Finally, the last one also employs ML but this time following the *bag-of-concepts* (BoC) approach common-sense affective knowledge. Each approach has been adapted properly to English and Spanish, depending on the resources available for each language.

The remainder of the paper is structured as follows. In Section 2, we describe the approaches proposed, as well as the tools and resources used in the implementation. The experiments performed and their evaluation and discussion are provided in Section 3. Finally, Section 4 concludes the paper, and outlines the future work.

2 Polarity Classification

The following sections explain our three different approaches submitted to the polarity classification subtask of RepLab 2013. We focus on the techniques, tools and resources employed for the design and implementation of each approach. Their main goal is to determine whether a tweet has positive, negative or neutral impact on the reputation of a given entity. These approaches were properly adapted each approach to English and Spanish but, as not all the required resources are available for both languages, our adaptations are no symmetric.

The preprocessing module, common to all our approaches, is explained in Section 2.1. The first one is unsupervised and it is described in Section 2.2. In Section 2.3 and Section 2.4 we explain the supervised approaches.

2.1 Preprocessing

Tweets are preprocessed before applying any model by following these common steps, for both English and Spanish languages:

1) *Cleansing*. All the words with non-standard characters are removed.

[3] `http://www.twitter.com`

2) *Tokenisation.* The text is first split into sentences using regular expressions.
3) *Lemmatisation.* For each sentence we extract the lemmas of its words. In English texts this sentence extraction is made using the MBLEM[4] lemmatiser, that combines a memory-based ML algorithm with a dictionary lookup. Freeling[5] [3] was the tool selected for extracting lemmas from sentences in Spanish. In order to obtain accurate lemmas a custom dictionary was created to replace common out-of-vocabulary (OOV) words, such as misspellings and informal lexical variants with their canonical version (e.g. `lol` → `laugh`; `q` → `que`).
4) *URL removal.* Each URL is substituted with a place-holder tag (`_URL_`).
5) *Twitter hashtag splitting.* Hashtags can contain sentiment-related information so we split them into independent words using a cost function based on word frequencies (e.g. `#WeHateVF` → `we hate VF`).
6) *Emoticon normalisation.* We follow the same approach found in [4] in order to replace emoticons with their textual equivalence (e.g. `xDDD` → `I am happy`).
7) *Named-entity detection.* Locations, people and temporal expressions are detected using a *maximum entropy tagger*, which was trained with the CONLL dataset [5].

2.2 Volvam Polarity 1: Unsupervised Lexicon-Based Model

Our first submitted run makes use of the fuzzy lexicons of *SentiStrength*[6] [6], in order to detect the most common informal terminology used in Twitter. These lexicons indicate not only if a term represents a positive or a negative opinion, but also an intensity score. The terms in these lexicons are English terms, so we manually translated them to obtain the corresponding Spanish lexicons. In addition, we extended the lexicons to allow the detection of modifiers that can invert (negation), increase or decrease the polarity score of each term. The polarity score of a text T is calculated by adding the lexicon scores of each term t inside that text:

$$polarityScore(T) = \sum_{t \in T} lexiconScore(t) * modifiersScore(t) \qquad (1)$$

where $lexiconScore(t)$ is the polarity score for the term t (range $[-4, 4]$) and $modifiersScore(t)$ is the score given to the term t by the modifiers of term t (range $[-1, 1]$). Finally, the polarity of that text is assigned depending on the polarity score obtained, using the following formula:

$$polarity(T) = \begin{cases} positive & \text{if } polarityScore(T) > 0 \\ neutral & \text{if } polarityScore(T) = 0 \\ negative & \text{if } polarityScore(T) < 0 \end{cases}$$

[4] `http://ilk.uvt.nl/mbma/`
[5] `http://nlp.lsi.upc.edu/freeling/`
[6] `http://sentistrength.wlv.ac.uk`

2.3 Volvam Polarity 2: Supervised Model combining Lexicons and Concepts

Our second submitted run uses a supervised ML model. The features used for this model are generated using the unsupervised model from Section 2.2:

- *TotalPolarity*. Total polarity obtained from the unsupervised model.
- *AvgSubjectivity*. Average subjectivity values extracted from the v2.0 polarity dataset [7].
- *CountPositive*. Number of positive words in the text.
- *CountNegative*. Number of negative words in the text.
- *CountNeutral*. Number of neutral words in the text.
- *SentenceTokens*. Tokens/sentence ratio.
- *TotalSubjectivity*. Total subjectivity value,
- *CountSubjective*. Number of words with subjectivity > 0.
- *CountProfanity*. Number of profanity words.
- *CountQuestions*. Number of sentences that are questions.
- *CountNonQuestions*. Number of sentences that are not questions.
- *CountNegated*. Number of negated sentences.
- *CountModPlus*. Number of augmentative modifiers.
- *CountModMinus*. Number of diminutive modifiers.

In addition, for the English texts, we added emotion-based features from *SenticNet* [8]. SenticNet consists on a lexicon containing four concept dimensions for each term: *pleasantness*, *attention*, *sensitivity* and *aptitude*. These concepts and its scores are used as additional features to build the ML model.

As the training dataset provided for this task was highly unbalanced, in terms of language and polarity labels, we followed a cross-corpus approach. As training set for the English language we used the *sentiment analysis training dataset* from *SemEval 2013* [9] and, for the Spanish language, the *TASS 2012* [10] training set. The classification model was built using the *Random Forests* [11] ensemble classifier on a subset of 6000 tweets.

2.4 Volvam Polarity 3: Supervised Model using Bag-of-Concepts

In our last submission we created different models for each language. For English, a Random Forest classifier was built using concept count vectors extracted from the provided RepLab training data. We followed the BoC approach using SenticNet common-sense affective knowledge. As we did not find an equivalent emotion-based model for Spanish, we followed a simpler *bag-of-words* approach using the lemmas of the terms in the text.

3 Evaluation

Our system was evaluated in terms of *accuracy* and F(R, S)[12], where R (*reliability*) is the precision of relations predicted by the system with respect to

actual relations in the gold standard and S (*sensitivity*) is the recall of relations predicted by the system with respect to the actual relations in the gold standard. A comparative of the obtained results are detailed in Table 1 (only the best run for each one of the other teams is displayed for informative purposes).

Method	**Accuracy**	**Reliability**	**Sensitivity**	**F(R, S)**
SZTE_NLP_polarity_6	**0.685**	**0.465**	0.345	**0.381**
popstar_polarity_5	0.638	0.433	0.339	0,373
Daedalus_polarity_3	0.438	0.312	0.397	0.341
Volvam_polarity_2	0.408	0.313	0.394	0.340
Volvam_polarity_1	0.389	0.302	**0.402**	0.336
NLP_IR_GROUP_UNED_polarity_1	0.578	0.333	0.309	0.316
lia_polarity_5	0.644	0.446	0.268	0.311
UAMCLYR_polarity_05	0.577	0.329	0.286	0.300
replab2013_UNED_ORM_polarity_1	0.587	0.316	0.290	0.298
Baseline	0.584	0.315	0.289	0.297
GAVKTH_polarity_2	0.263	0.371	0.213	0.267
Volvam_polarity_3	0.537	0.315	0.225	0.255
diue_polarity_1	0.546	0.333	0.215	0.254
IE-Polarity-4	0.513	0.279	0.222	0.212
ALLPOSITIVE	0.577	1	0	0

Table 1. Polarity classification results at RepLab 2013.

In general, the results obtained by all participants are not as high as the state-of-the-art results in polarity classification. This happens because polarity for reputation is a more complex task [1]. In addition, the datasets provided are highly unbalanced so the accuracies are no significant [13, 14]. This fact can be seen in the results of the trivial `ALLPOSITIVE` run, where all texts in the training set were classified as positive, which achieves an accuracy of 57%.

Our best ranked approach is the second one with a F-score of 34%, very near to the 38% obtained by the best approach of all participants. Our first approach reached the best sensitivity of all runs, with a 40%.

4 Conclusions

In this paper we described our participation in the profiling (polarity classification) task of the RepLab 2013 workshop. We covered three different approaches, one unsupervised and two unsupervised, combining machine learning and lexicon-based techniques with an emotional concept model. These approaches were properly adapted to English and Spanish depending on the resources available for each language. We obtained promising results in the overall evaluations, reaching a F-score of 34% and a sensitivity of 40% in the best cases. The reasonable level of performance compared to other methods encourages us to continue working on the improvement of the proposed approaches.

References

1. Balahur, A.: The challenge of processing opinions in online contents in the social web era. In: Proceedings of the Language Engineering for Online Reputation Management Worksop, LREC 2012. (2012)
2. Amig, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martn, T., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2013: Evaluating online reputation monitoring systems. In: Fourth International Conference of the CLEF initiative, CLEF 2013, Valencia, Spain. Proceedings. Springer LNCS (2013)
3. Padr, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA) (may 2012)
4. Mosquera, A., Lloret, E., Moreda, P.: Towards facilitating the accessibility of web 2.0 texts through text normalisation. In: Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA) ; Istanbul, Turkey. (2012) 9–14
5. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics (2003) 142–147
6. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology **61**(12) (2010) 2544–2558
7. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL. (2004)
8. Cambria, E., Havasi, C., Hussain, A.: Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In Youngblood, G.M., McCarthy, P.M., eds.: FLAIRS Conference, AAAI Press (2012)
9. Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., Ritter, A.: Semeval-2013 task 2: Sentiment analysis in twitter. In: Proceedings of the International Workshop on Semantic Evaluation, SemEval. Volume 13. (2013)
10. Villena-Román, J., Lana-Serrano, S., Martínez-Cámara, E., Cristóbal, J.C.G.: Tass - workshop on sentiment analysis at sepln. Procesamiento del Lenguaje Natural **50** (2013) 37–44
11. Breiman, L.: Random forests. Mach. Learn. **45**(1) (October 2001) 5–32
12. Amigo, E., Gonzalo, J., Verdejo, F.: Reliability and sensitivity: Generic evaluation measures for document organization tasks. In: Tech. rep., UNED. (2012)
13. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1999) 42–49
14. Boldrini, E., Fernández Martínez, J., Gómez Soriano, J.M., Martínez Barco, P., et al.: Machine learning techniques for automatic opinion detection in non-traditional textual genres. (2009)