Sentiment Analysis in Twitter for Spanish

Ferran Pla and Lluís-F. Hurtado

Universitat Politècnica de València
Camí de Vera s/n, 46022 València
{fpla,lhurtado}@dsic.upv.es

Abstract. This paper describes a SVM-approach for Sentiment Analysis (SA) in Twitter for Spanish. This task was part of the TASS2013 workshop, which is a framework for SA that is focused on the Spanish language. We describe the approach used, and we present an experimental comparison of the approaches presented by the different teams that took part in the competition. We also describe the improvements that were added to our system after our participation in the competition. With these improvements, we obtained an accuracy of 62.88% and 70.25% on the SA test set for *5-level* and *3-level* tasks respectively. To our knowledge, these results are the best results published until now for the SA tasks of the TASS2013 workshop.

Keywords: Sentiment Analysis, Twitter, Machine Learning.

1 Introduction

Twitter has become a popular micro-blogging site in which users express their opinions on a variety of topics in real time. The nature of texts used in Twitter (ungrammatical sentences with a lot of emoticons, abbreviations, specific terminology, slang, etc.) poses new challenges for researchers in Natural Language Processing (NLP) to provide effective solutions for Sentiment Analysis (SA) in micro-blogging texts. Therefore, the usual techniques of NLP must be adapted to these characteristics of the language, and new approaches must be proposed in order to successfully address this problem. NLP tools like POS taggers, parsers, or Named Entity Recognition (NER) tools usually fail when processing tweets because they generally are trained on grammatical texts and they perform poorly in micro-blogging texts.

Most of the work of SA on Twitter is for the English language and this is also true for the resources and tools available for NLP. Therefore, the TASS2013 workshop aims to be a framework for SA and on-line reputation analysis that is focused on the Spanish language. The organization of TASS2013 provided a corpus of Spanish tweets, *The General Corpus* [13], which is annotated with their polarity. This is a very important resource that allows researchers to compare their approaches for the SA problem on Twitter by using the same data.

SA has been widely studied in the last decade in multiple domains. Most work focuses on classifying the polarity of the texts as positive, negative, mixed,

or neutral. The pioneering works in this field used supervised [7] or unsupervised (knowledge-based) [12] approaches. In [7], the performance of different classifiers on movie reviews was evaluated. In [12], some patterns containing POS information were used to identify subjective sentences in reviews to then estimate their semantic orientation. The construction of polarity lexicons is another widely explored field of research. Opinion lexicons have been obtained for English [3] [15] and also for Spanish [8].

Research works about SA on Twitter are much more recent. Twitter appeared in the year 2006 and the early works in this field are from 2009 when Twitter started to achieve popularity. Some of the most significant works are [1], [2], and [5]. A survey of the most relevant approaches to SA on Twitter can be see in [4] and [14]. The SemEval2013 competition has also dedicated a specific task for SA on Twitter [16], which shows the great interest of the scientific community.

In this work we present our approach and the results obtained for the SA tasks proposed at the TASS2013 workshop. Two different sub-tasks called *5-level* and *3-level* were proposed. Both sub-tasks differ only in the polarity granularity considered. The *5-level* sub-task uses *N* and *N+* labels for negative polarity, *P* and *P+* labels for positive polarity, and the *NEU* label for neutral polarity. The *3-level* sub-task only has three polarity levels: *N*, *P*, and *NEU*. In both sub-tasks, an additional label (*NONE*) was used to represent tweets with no polarity at all (objective tweets). The data used at TASS2013 workshop contains approximately 68000 Twitter messages (*tweets*) written in Spanish (between November 2011 and March 2012) by about 150 well-known personalities of the world of politics, economy, communication, mass media, and culture. Each tweet includes its ID, the creation date, and the user ID. The corpus is encoded in XML and it is divided into two sets: training (about 10%, 7219 tweets) and test (about 90%, 60798 tweets). The distributions per polarity of the training set is: 18.49% for N, 11.73% for N+, 9.28% for NEU, 20.54% for NONE, 17.07% for P, and 22.88% for P+.

2 System Description

The SA system proposed consists of 3 modules. The first module is the Pre-processing module, which performs the tokenization, lemmatization, and NER of the input tweet. A lemma reduction and a POS tagging process is also carried out in this module. The second module is the Feature Extraction module, which selects the features from the pre-processed tweet and obtains a feature vector. Some features require the use of a polarity lexicon of lemmas and words. To determine the best features, a tuning process is required during the training phase. The third module is the Polarity Classifier module, which uses a classifier to assign a polarity label to the tweet.

Before addressing the SA tasks, it is necessary to make a proper tokenization of the tweets that make up the training corpus and test corpus. Although there are a lot of tokenizers available on the web, they need to be adapted in order to address the tokenization of a tweet. In our system, we decided to use and

adapt available tools for tokenization, lemmatization, NER, and POS tagging. We adapted the package *Tweetmotif* that is described in [5] to process Spanish tweets. We also used *Freeling* [6] (with the appropriate modifications for handling Twitter messages) for stemming, NER, and POS tagging. We added some functions to group special tokens into sigle tokens (e.g., *hashtags*, *web* addresses, *url*, *dates*, *numbers*, and some *punctuation* marks).

The SA task was addressed as a classification problem that consisted of determining the polarity of each tweet. We used WEKA, which is a tool that includes (among other utilities) a collection of machine-learning algorithms that can be used for classification tasks. Specifically, we used a SVM-based approach because it is a well-founded formalism, that has been successfully used in many classification problems. In the SA task, SVM has shown it ability to handle large feature spaces and to determine the relevant features. We used the NU-SVM algorithm [11] from an external library called *LibSVM*, which is very efficient software for building SVM classifiers. It is easy to integrate this software with WEKA thus allowing us to use all of WEKA's features. We used the *bag_of_words* approach to represent each tweet as a feature vector that contains the occurrences of the selected features.

The tuning process carried out had two objectives: to choose which features to include in the model, and to perform the parameter estimation of the SVM. We conducted this optimization by means of a 10-fold cross validation using the official TASS2013 training set as the development set. In order to determine the features of the model, the following four parameters were considered: *lemma frequency (f)*, *bilemma*, *selPOS*, and *polarity lexicon (DIC)*. The *lemma frequency (f)* parameter determines the minimum frequency necessary to consider a lemma as a feature. The *bilemma* parameter determines if bigrams of lemmas (in addition to single lemmas) are included as features in the model. The *selPOS* parameter determines if only the lemmas that belong to a prefixed set of POS are included in the model. When *selPOS* is used, only those lemmas belonging to *nouns, verbs, adjectives, adverbs* POS (in addition to the *emoticons* and *exclamations*) are included in the model. Finally, *DIC* determines if external polarity lexicons are used. The external lexicons used by our system in the TASS2013 competition [9] were lists of words and lemmas with their a priori polarity. One of the lexicons used was originally for English [15] that was translated into Spanish automatically, and other [8] lexicon was a list of words that was originally in Spanish. With these two resources we constructed our original dictionary (*DIC*). Then, we combined *DIC* with the lexicon presented in [10] in order to obtain an improved lexicon (*DIC-improved*).

Table 1 shows the *Accuracy* and confidence interval (with a 95% level of confidence) of the 10-fold cross validation process for *5-level* and *3-level* subtasks and for different combinations of the features under consideration (from system s_1 to system s_{14}). We also include the average number of features for each system.

The *Accuracy* results obtained by the different systems considered were not statistically significant in many cases. However, there was a great difference in

performance between the systems that did not use lexicons and those that did use them; especially those using the *DIC-improved* lexicon. A detailed analysis of the hits obtained by the systems showed that there was a different of up to 5% for the correctly labeled tweets even on systems with the same precision.

Table 1. *Accuracy* results for the tuning process using the training set (10-fold).

System	Features	# features	Accuracy (%)	
			5-level	3-level
s_1	f=1	11436.7	45.41 ± 1.14	62.35 ± 1.33
s_2	f=1+selPOS	11308.7	44.99 ± 1.32	60.58 ± 1.17
s_3	f=1+DIC	11438.7	46.74 ± 1.06	65.16 ± 1.43
s_4	f=1+selPOS+DIC	11310.7	46.38 ± 1.05	64.37 ± 1.22
s_5	f=1+DIC-improved	11438.7	49.84 ± 1.23	68.17 ± 1.44
s_6	f=1+bilemma+DIC-improved	64686.7	49.63 ± 1.01	67.53 ± 1.08
s_7	f=1+selPOS+DIC-improved	11310.7	**50.20 ± 1.55**	67.20 ± 1.26
s_8	f=2	4533.0	45.93 ± 1.16	62.47 ± 1.07
s_9	f=2+DIC-improved	4535.0	50.12 ± 1.24	**68.35 ± 1.28**
s_{10}	f=2+bilemma+DIC-improved	16153.3	49.61 ± 1.37	68.04 ± 0.91
s_{11}	f=2+selPOS+DIC-improved	4410.0	50.09 ± 1.27	67.47 ± 1.30
s_{12}	f=3+DIC-improved	3015.3	49.85 ± 1.90	67.95 ± 1.41
s_{13}	f=3+bilemma+DIC-improved	9049.0	49.40 ± 1.29	67.78 ± 1.08
s_{14}	f=3+selPOS+DIC-improved	2904.3	49.73 ± 1.53	67.16 ± 1.43

Taking this into account, we decided to combine the systems in order to take advantage of their complementarity. Several different combination methods were tested and no relevant differences in accuracy were found. Finally, we decided to use a majority voting scheme. Each tweet was classified by each system and the polarity that was chosen by the majority of the systems was the polarity definitively assigned to the tweet. If a tie occurred, the most frequent among the tied polarities (in the training set) was selected. In the experimental work conducted, all the possible combinations of systems were tested.

The best results was obtained by combining 2 systems. When we used the system *voting1* (by combining s_7 and s_{13} systems) we improved the accuracy from 50.20% to 50.45% for the *5-level* task. With the system *voting2* (by combining s_{11} and s_{13} systems) we improved from 68.35% to 68.68% for *3-level* task.

3 The Evaluation on the Test Set

A total of 13 teams participated in the TASS2013 SA task. Fifty-six runs were submitted for evaluation in the competition. The official results ranged from 61.6% to 13.5% (for *5-level* task) and from 66.3% to 38.8% (for *3-level* task). The best results were obtained by machine learning-based approaches. A detailed description of the different approaches is available on the TASS2013 website.

We constructed new models for the *5-level* and *3-level* tasks with the best set of features obtained in the tuning phase. We tested these models on the

test set supplied at the TASS2003 competition. The results obtained with the confidence interval are show in Table 2. It also include the 3 best approaches at the TASSS2013 competition: UA, ELHUYAR and UPV-ELiRF(our system).

Table 2. *Accuracy* results on the test set.

System	Accuracy (%)	
	5-level	3-level
s_7	**60.02 ± 0.39**	68.85 ± 0.37
s_9	59.21 ± 0.39	**69.64 ± 0.37**
voting1	**62.88 ± 0.38**	70.16 ± 0.36
voting2	62.77 ± 0.38	**70.25 ± 0.36**
UA	61.62 ± 0.39	66.28 ± 0.38
ELHUYAR	60.10 ± 0.39	68.65 ± 0.37
UPV-ELiRF	57.60 ± 0.39	67.40 ± 0.37

Note that *Accuracy* results are higher than those obtained on the training set. This was true for all of the approaches presented at this competition. We have no clear explanation for this, it may be because the distribution of tweets by category in the training set (i.e, P+,22%; NONE,20%) is different from the test set (i.e, P+,34%; NONE,35%), or it may be because the process of manual supervision was different for these training and test sets. Our best voting systems outperform the ELHUYAR and UA systems for both the *3-level* and the *5-level* tasks with statistical significance. For the *3-level* task, our individual system s_9 also outperformed the other approaches with statistical significance.

4 Conclusions

In this paper, we have presented our approach for the SA task of the TASS2013 competition. For the classification stage, we used a Support Vector Machine approach with WEKA and the external LibSVM library.

We have presented the improvements we have made to the system that we submitted to the TASS2013 competition. These improvements consisted of adding new features to the classifiers, the construction of new polarity dictionaries, and the combination of different models by means of voting techniques. With these improvements, we obtained the best results for *5-level* and *3-level* tasks with an accuracy of 62.88% and 70.25% respectively. We think that the corpus and the gold standards provided at the TASS2013 competition (which are available on the TASS2013 webpage) and the evaluation presented in this work will be helpful for other research groups that are interested in the SA task.

As future work, we plan to continue working on this task, taking into account new features and resources. Specifically, these can include using more accurate text normalization techniques for improving POS tagging and NER for tweet domain and using and adapting new language resources for conducting a deep syntactic analysis to tackle some specific issues that could improve SA tasks in Twitter, such as negation, modifiers of polarity (adverbs), or coreference.

Acknowledgments. This work has been funded by the projects, DIANA (MEC TIN2012-38603-C02-01) and Tímpano (MEC TIN2011-28169-C05-01).

References

1. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics (2010) 36–44
2. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology **60**(11) (2009) 2169–2188
3. Liu, B., Hu, M., Cheng, J.: Opinion observer: Analyzing and comparing opinions on the web. In: Proceedings of the 14th International Conference on World Wide Web. WWW '05, New York, NY, USA, ACM (2005) 342–351
4. Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A., Montejo-Raéz, A.: Sentiment analysis in twitter. Natural Language Engineering **1**(1) (2012) 1–28
5. O'Connor, B., Krieger, M., Ahn, D.: Tweetmotif: Exploratory search and topic summarization for twitter. In Cohen, W.W., Gosling, S., eds.: Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010, The AAAI Press (2010)
6. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, ELRA (May 2012)
7. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: IN PROCEEDINGS OF EMNLP. (2002) 79–86
8. Perez-Rosas, V., Banea, C., Mihalcea, R.: Learning sentiment lexicons in spanish. In Chair), N.C.C., Choukri, K., Declerck, T., Doğan, 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)
9. Pla, F., Hurtado, L.F.: Tass-2013: Análisis de sentimientos en twitter. In: Proceedings of the TASS workshop at SEPLN 2013, IV Congreso Español de Informática (2013)
10. Saralegi, X., San Vicente, I.: Elhuyar at tass 2013. In: Proceedings of the TASS workshop at SEPLN 2013, IV Congreso Español de Informática (2013)
11. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Comput. **12**(5) (May 2000) 1207–1245
12. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: ACL. (2002) 417–424
13. Villena-Román, J., García-Morera, J.: Workshop on sentiment analysis at sepln 2013: An over view. In: Proceedings of the TASS workshop at SEPLN 2013, IV Congreso Español de Informática (2013)
14. Vinodhini, G., Chandrasekaran, R.: Sentiment analysis and opinion mining: A survey. International Journal **2**(6) (2012)
15. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: Opinionfinder: A system for subjectivity analysis. In: Proceedings of HLT/EMNLP on Interactive Demonstrations, Association for Computational Linguistics (2005) 34–35
16. Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., Ritter, A.: Semeval-2013 task 2: Sentiment analysis in twitter. Proceedings of the International Workshop on Semantic Evaluation, SemEval **13** (2013)