Experiments on feature replacements for polarity classification of Spanish tweets

Experimentos sobre sustituciones de características para la clasificación de la polaridad de tweets en español

José M. Perea-Ortega **Alexandra Balahur**

European Commission, Joint Research Centre (JRC)
Institute for the Protection and Security of the Citizen (IPSC)
Via Fermi 2749, 21027 Ispra (VA), Italy

{jose-manuel.perea-ortega,alexandra.balahur}@jrc.ec.europa.eu

Resumen: En este artículo presentamos varios experimentos para abordar la tarea de la clasificación global de la polaridad de tweets en español. Estos experimentos se han centrado en diferentes sustituciones de características llevadas a cabo para ambos conjuntos de datos proporcionados, tanto de desarrollo como de test. Las sustituciones realizadas se basaron principalmente en los signos de puntuación repetidos, los emoticonos y las palabras de opinión, mediante el uso de un diccionario construido para análisis de sentimientos. A continuación, se aplicó un enfoque basado en aprendizaje automático para obtener la polaridad de los tweets. Los resultados obtenidos muestran que las estrategias hibridas propuestas mejoran la precisión en la clasificación de la polaridad de los sentimientos expresados en tweets en comparación con el uso de caracteristicas basadas solo en n-gramas.
Palabras clave: Análisis de sentimientos, Clasificación de la polaridad, Aprendizaje automático, Twitter

Abstract: In this paper we present several experiments to address the global polarity classification task of Spanish tweets. These experiments have focused on different feature replacements carried out for both the development and test data sets provided. The replacements performed were mainly based on repeated punctuation signs, emoticons and affect words, by using an in-house built dictionary for sentiment analysis. Then, a machine learning approach was applied to get the polarity of the tweets. The results obtained show that the hybrid approaches proposed improve sentiment polarity classification when compared to simple n-gram feature use.
Keywords: Sentiment Analysis, Polarity Classification, Machine Learning, Twitter

1. Introduction

In the past years, we have witnessed an unprecedented growth in the quantity of information that is being produced and shared on the Internet. Using Social Media platforms such as Twitter, Facebook, Instagram, etc. people obtain, share and comment the information related to events, persons, organizations in almost real time. As such, this highly dynamic information, if filtered appropriately, can be employed to obtain an accurate snapshot of the current state of events, as well as people's attitudes towards them.

This valuable knowledge, as it has been shown by the scientific literature of the past decade, can be of real benefit to many types of applications, in Economics, Marketing, e-Law Making, detecting public take on critical aspects of society, to mention but a few. Nonetheless, the path from the information to the knowledge is not a straightforward one. Extracting knowledge from the large quantities of information produced requires the use of automatic opinion processing tools.

The approaches presented in this paper deal with automatic opinion detection and classification from social media texts, more specifically, Twitter. The main idea is to build hybrid supervised models, in which the knowledge extracted from lexical resources of affective terms can be exploited to increase the accuracy of the sentiment classification. We experiment with different combinations of features such as unigrams, bigrams, skip-grams and mixtures thereof, lexicons and features specifically designed to take into account the special language contained in social media, e.g. emoticons, hashtags, punctuation signs. Our extensive experiments show that the application of the proposed techniques should be taken into account in polarity classification systems and specifically those related to the emoticon replacements.

The rest of the paper is organized as follows: Section 2 describes research that deals with sentiment analysis in short informal texts originating from social media sites; Section 3 describes the framework and the data sets used as well as the data processing carried out; the proposed feature replacements are presented in Section 4; Section 5 describes the different experiments performed and the results obtained; finally, conclusions and further work are expounded in Section 6.

2. Related work

Read (2005), Go, Bhayani, and Huang (2009) and Pak and Paroubek (2010) first construct a corpus and subsequently evaluate a method to classify sentiment in tweets based on the presence of emoticons (e.g. ":)", ":(", etc.) as markers of positive and negative tweets. Zhang et al. (2011) employ a hybrid approach, combining supervised learning with the knowledge on sentiment-bearing words, which they extract from the DAL sentiment dictionary (Whissell, 1989). Their pre-processing stage includes the removal of retweets, translation of abbreviations into original terms and deleting of links, a tokenization process, and part-of-speech tagging. They employ various supervised learning algorithms to classify tweets into positive and negative, using n-gram features with SVM and syntactic features with Partial Tree Kernels, combined with the knowledge on the polarity of the words appearing in the tweets.

In the 2013 edition of TASS (Villena-Román et al., 2014), most systems employed a supervised approach using n-gram features enriched with semantic knowledge from sentiment dictionaries or linguistic processing tools (e.g. FreeLing, Twitmotif[1]). Remaining ones employed approaches based on averaging sentiment values associated to terms extracted from affect dictionaries or, in a more complex setting, computing overall values using sentiment dictionaries in conjunction to graph-based methods.

In the TASS 2012 evaluation campaign the best participating system employed 5 classifiers to distinguish among the 5 classes of polarity (among themselves) and the objective class (separately) (Villena-Román et al., 2013). SemEval 2013 and SemEval 2014 also contained tasks on sentiment analysis in tweets (Wilson et al., 2013). In both editions of this evaluation campaign, the best-performing systems used additional dictionaries that were built from large data sets and word-emotion association dictionaries built from millions of tweets gathered based on specifically designed affect-based patterns.

3. Framework

Different tools and resources have been used to perform the experiments for the second participation in the TASS workshop. Since our approach is based on machine learning, WEKA[2] has been used as a tool for generating the different learning models, by applying Support Vector Machines Sequential Minimal Optimization (SVM SMO) (Platt, 1998) as learning algorithm. SVM has been proven to be highly effective in traditional text categorization and has been applied successfully in many opinion mining tasks overcoming other machine learning techniques (O'Keefe and Koprinska, 2009; Esuli and Sebastiani, 2005).

3.1. Data processing

The only data sets used to carry out the experiments have been those provided by TASS for the *sentiment analysis at global level* task[3]. The *general corpus* contains over 68,000 tweets in Spanish about well-known personalities in politics, economics, communication or culture, and they were collected between November 2011 and March 2012.

[1] https://github.com/brendano/tweetmotif

[2] http://www.cs.waikato.ac.nz/ml/weka

[3] http://www.daedalus.es/TASS2014/tass2014.php#corpus

This corpus was divided into two sets: training (about 10 %) and test (90 %). The training set contains 7,219 tweets and they were tagged with global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. Five sentiment levels have been defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE).

The basic processing carried out for both data sets (training and test) consisted of removing the URLs and numbers as well as replacing two or more occurrences of vowels and symbols like '(' or ')' by only one occurrence. Thus, emoticons like ';-))))' were replaced by ';-)' or tokens like 'jaaaajaa' or 'largoooooo' were replaced by 'jaja' and 'largo' respectively. However, we kept the main token from the users and *hashtags* without the symbols '@' and '#'. No stemming process was applied and stop words were removed for some experiments. We used a *light* version of the stop word list provided by Snowball[4] for Spanish language, by manually discarding some words that might have influence in the polarity, such as *no*, *sí*, *más*, *mucho*... Finally, from the total of 325 stop words included in the original Snowball list, 228 were removed, remaining then a final list of 97 stop words.

3.2. Semantic resources

Several semantic resources have been used in order to apply our approach based on feature replacements. We have generated a preliminary version of a new semantic resource called **USELESP** (Unified SEntiment LExicon for SPanish). Our idea is to integrate and unify existing semantic resources in different languages into an unique Spanish lexicon for sentiment analysis. For the first version of USELESP two main resources have been integrated:

- **Spanish JRC sentiment dictionary** (Steinberger et al., 2011). This lexicon was generated by gathering and filtering English sentiment terms from the resources Micro-WordNet (Cerini et al., 2007) and JRC Tonality Dictionary (Balahur et al., 2009). The resources were then translated to Spanish by Google translator and filtered afterward. In addition, they were mapped to four categories (positive, highly positive, negative and highly negative) in order to distinguish two levels of intensity. The JRC sentiment dictionary contains a total of 1.638 terms, of which 466 are highly negative, 550 negative, 503 positive and 119 highly positive.
- **eSOL lexicon** (enriched Spanish Opinion Lexicon) (Molina-González et al., 2013). This resource was generated by machine translation of the sentiment words provided by the Bing Liu English Lexicon (Hu and Liu, 2004) and by following a corpus-based approach. Reverso was used as machine translation tool and MuchoCine (Cruz et al., 2008) as corpus of film reviews. Finally, all the sentiment words were manually revised. The lexicon is composed of 2.536 positive words and 5.639 negative ones.

The JRC sentiment dictionary includes some terms ending with the symbol ' %', which means that different morphological variants can be derived from such term. Therefore, during the generation of USELESP, we created all the possible derivations for those terms and checked them against the Spanish dictionary provided by Freeling[5] (Padró and Stanilovsky, 2012). Then, the existing derivations were added as new sentiment words into USELESP.

Finally we have also used the emoticon list provided by SentiStrength[6] (Thelwall et al., 2010). This list is composed of around 106 emoticons with an associated sentimental weight (1 for positive emoticons and -1 for negative ones).

4. Feature replacements

Our main goal in TASS 2014 was to test the performance of different feature replacements which had been previously proved with success in the "Sentiment Analysis in Twitter" task of SemEval 2013 for English (Balahur, 2013). The proposal behind this approach is to put under the same label different features which have the same sentiment. For other cases, it is simply to replace specific features commonly used in tweets like repeated punctuation signs by the same label. Thus, the

[4] http://snowball.tartarus.org/algorithms/spanish/stop.txt

[5] http://nlp.lsi.upc.edu/freeling

[6] http://sentistrength.wlv.ac.uk

Skip-gram model	Content
Bi-grams	*desempleo desempleo_eeuu eeuu eeuu_baja baja baja_niveles niveles niveles_marzo marzo*
1-skip-bi-grams	*desempleo desempleo_baja desempleo_eeuu eeuu eeuu_niveles eeuu_baja baja baja_marzo baja_niveles niveles niveles_marzo marzo*
2-skip-bi-grams	*desempleo desempleo_niveles desempleo_baja desempleo_eeuu eeuu eeuu_marzo eeuu_niveles eeuu_baja baja baja_marzo baja_niveles niveles niveles_marzo marzo*

Table 1: Skip-grams examples generated for the original tweet "*El desempleo en EEUU baja a los niveles de marzo de 2009*" without stop words and numbers

total number of features during the learning process would be reduced. The replacements carried out were the following:

- **RPSR** (Repeated Punctuation Signs Replacement). Repetitions of punctuation signs such as '.', '!' or '?' were replaced by fixed labels like *multipoint*, *exclamation* or *question*, respectively.
- **ER** (Emoticon Replacement). The emoticons found in the content of the tweets were replaced by the fixed labels *positive* or *negative* depending on the sentimental weight assigned by the emoticon list used as semantic resource (see previous section).
- **AWR** (Affect Word Replacement). The sentiment words found in the content of the tweets were replaced by the fixed labels *hpositive*, *positive*, *hnegative* or *negative* according to the category assigned by USELESP, the new semantic lexicon generated from several resources as explained above.

In addition to the feature replacements, the use of **skip-grams** was also examined. Skip-grams are a technique largely used in the field of speech processing, in which not only adjacent sequences of words are formed (bi-grams, tri-grams, etc.) but it also allows to skip tokens in between (1-skip-bi-grams, 2-skip-tri-grams, etc.) (Guthrie et al., 2006). Table 1 shows several examples of the skip-grams generated for the original tweet "*El desempleo en EEUU baja a los niveles de marzo de 2009*" without stop words and numbers.

5. Experiments and results

According to the feature replacements explained above, many experiments were performed by combining the different types of replacement and skip-grams models. Nevertheless, the highest results in terms of accuracy were always obtained by the joint application of uni-grams and bi-grams, instead of using uni-grams or bi-grams separately. Taking into account the id labels used for each feature replacement explained in Section 4 (RPSR, ER and AWR) and the n-skip-grams models implemented, Table 2 shows the experiments that obtained highest accuracies. The accuracy score was calculated by considering six target categories (P+, P, N+, N, NONE and NEU). For all the experiments a basic data processing was performed and SVM SMO was applied as learning algorithm, as explained above in Section 3.

From the results shown in Table 2 we can see that there are no significant improvements between the different types of replacements carried out, since all of them are around 0.48 or 0.47 of accuracy. The highest score (0.4838) was obtained by applying the emoticon replacement solely with bi-grams. However, it is noteworthy the improvement achieved by the best experiment regarding the baseline one, with a difference of +2.18 %.

6. Conclusions and further work

This paper presents the experiments and results obtained by the JRC team in the *sentiment analysis at global level task* in TASS 2014. In our second participation, we continue applying an approach based on machine learning but, as novelty, we proposed several techniques of feature replacements and the use of skip-grams. Three types of features were chosen to be replaced by fixed labels: repeated punctuation signs, emoticons and sentiment words. Moreover, a preliminary version of a new semantic resource called USELESP (Unified SEntiment LExicon

Experiment ID	Description	Acc.
ER	emoticon replacement with bigrams	**0.4838**
RPSR-ER-AWR	all the feature replacements combined with bigrams	0.4823
ER-AWR	emoticon and affect word replacements with bigrams	0.4804
RPSR-ER-AWR-2-skip-bigrams	all the feature replacements combined with 2-skip-bigrams model	0.4784
baseline-stop	no replacements and no stop words with bigrams	0.4771
AWR	affect word replacement with bigrams	0.4770
AWR-2-skip-bigrams	affect word replacement with 2-skip-bigrams model	0.4761
RPSR-ER-AWR-1-skip-bigrams	all the feature replacements with 1-skip-bigrams model	0.4760
RPSR-AWR	repeated punctuation signs and affect word replacements with bigrams	0.4737
baseline	no replacements with bigrams	0.4735

Table 2: Experiments and results obtained for the *5-level sentiment analysis at global level task* in TASS 2014

for SPanish) was generated in order to detect and classify the sentiment words. Although the results obtained were not very encouraging, several conclusions were inferred from the experiments carried out:

- The application of the feature replacement techniques might be worth regarding the non-application of any replacement. It was shown the improvement obtained by applying the emoticon replacement technique regarding the baseline case (+2.18 % of accuracy).
- From the proposed feature replacements, there was no one that showed a significant improvement regarding the others. Therefore, any of the proposed techniques might be useful for polarity classification.
- The use of bi-grams performed better than any other n-grams model.
- The skip-grams models did not work as well as expected. The main reason might be that we applied them without removing the stop words, so it might be adding noise in the training data.

For further work we would like to improve the preliminary version of USELESP, by adding and manually reviewing new existing semantic resources in different languages. Besides, a deeper analysis of the behavior of the skip-grams models should be carried out in order to understand why they did not work as expected. Finally, we would like to explore the performance of the *word embedding* technique for short texts.

References

Balahur, A. 2013. OPTWIMA: Comparing knowledge-rich and knowledge-poor approaches for sentiment analysis in short informal texts. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 460–465. ACL.

Balahur, A., R. Steinberger, E. Van der Goot, B. Pouliquen, and M. Kabadjov. 2009. Opinion Mining on Newspaper Quotations. In *Proceedings of the workshop 'Intelligent Analysis and Processing of Web News Content' (IAPWNC), held at the 2009 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Milano, Italy.*

Cerini, S., V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini, 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 re-

sources for opinion mining. Franco Angeli Editore, Milano, IT.

Cruz, F., J.A. Troyano, F. Enríquez, and J. Ortega. 2008. Experiments in sentiment classification of movie reviews in Spanish. *Procesamiento del Lenguaje Natural*, 41:73–80.

Esuli, Andrea and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM-05)*, pages 617–624.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

Guthrie, D., B. Allison, W. Liu, L. Guthrie, and Y. Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the Fifth international Conference on Language Resources and Evaluation (LREC-2006)*.

Hu, M. and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177. ACM.

Molina-González, M. Dolores, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, and José M. Perea-Ortega. 2013. Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40:7250–7257.

O'Keefe, Tim and Irena Koprinska. 2009. Feature Selection and Weighting Methods in Sentiment Analysis. In *Proceedings of 14th Australasian Document Computing Symposium*.

Padró, Ll. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *LREC*, pages 2473–2479. European Language Resources Association (ELRA).

Pak, Alexander and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 19–21.

Platt, John C. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances in Kernel Methods - Support Vector Learning.

Read, Jonathon. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48.

Steinberger, J., P. Lenkova, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, R. Steinberger, H. Tanev, V. Zavarella, and S. Vázquez. 2011. Creating sentiment dictionaries via triangulation. In *Proceedings of WASSA 2011*, WASSA '11, pages 28–36. ACL.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. 2010. Sentiment in short strength detection informal text. *JASIST*, 61(12):2544–2558.

Villena-Román, Julio, Janine García-Morera, Sara Lana-Serrano, and José Carlos González Cristóbal. 2014. TASS 2013 - A Second Step in Reputation Analysis in Spanish. *Procesamiento del Lenguaje Natural*, 52:37–44.

Villena-Román, Julio, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González Cristóbal. 2013. TASS - Workshop on Sentiment Analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50:37–44.

Whissell, Cynthia. 1989. The Dictionary of Affect in Language. In *Emotion: theory, research and experience*, volume 4, The measurement of emotions. Academic Press, London.

Wilson, Theresa, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. SemEval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, June.

Zhang, Ley, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical Report HPL-2011-89, HP, 21/06/2011.