Experiments using varying sizes and machine translated data for sentiment analysis in Twitter

Experimentos utilizando diferentes tamaños y datos automáticamente traducidos para análisis de sentimientos en Twitter

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

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

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

Resumen: En este artículo presentamos varios experimentos para la tarea de análisis de sentimientos a nivel global dentro de la campaña de evaluación TASS. El objetivo de esta tarea es evaluar la polaridad global de textos cortos en español extraídos de Twitter. Para abordar esta tarea se ha aplicado un enfoque basado en aprendizaje automático probando diferentes combinaciones de características. Se han empleado varios diccionarios y un corpus traducido automáticamente para entrenamiento, adaptando al español un enfoque inicial diseñado para trabajar con textos en inglés. Además, se probaron en cascada cuatro clasificadores separados para determinar el sentimiento desde clases de polaridad más generales a más precisas. Aunque ésta es nuestra primera participación, los enfoques propuestos se podrían considerar buenas estrategias para generar corpus de entrenamiento para sistemas de clasificación de la polaridad en español
Palabras clave: Análisis de sentimientos, Clasificación de la polaridad, Aprendizaje automático, Twitter

Abstract: In this paper we present several experiments for the task entitled sentiment analysis at global level within the TASS evaluation campaign. The aim of this task is to assess the global polarity of Spanish short texts extracted from Twitter. To tackle this task, an approach based on machine learning by trying different feature combinations was applied. Several in-house built dictionaries and machine-translated data for training were employed by adapting an approach designed for English to Spanish. Additionally, four separate classifiers were tested in cascade to determine the sentiment from the general to the finer-grained classes of polarity. Although this is our first participation, the proposed approaches might be considered good strategies to generate learning data for polarity classification systems in Spanish
Keywords: Sentiment Analysis, Polarity Classification, Machine Learning, Twitter

1. Introduction

In the past decade, the quantity of user-generated contents on the Internet has been growing exponentially. Social Media platforms, such as Facebook, Twitter, Flickr, LinkedIn, etc., as well as commercial sites, like Amazon, Booking.com, etc. offer their users the possibility to share their experiences and opinions of topics ranging from economics, to politics, products, VIPs and globally-critical events. The value of such unbiased, real-time user-generated content has been shown to be tremendous, with applications in Marketing, Decision Support Systems, Politics and Public Policy support, disaster and crisis management, etc. Since the high volume of opinionated information makes its manual processing virtually impossible, systems have been developed to treat texts and process the opinions they contain automatically, in the context of the Subjectivity and Sentiment Analysis tasks, within the field of Natural Language Processing (NLP).

Subjectivity and Sentiment Analysis typically aim at detecting subjective, "private" states (i.e. opinions, emotions, sentiments, evaluations, beliefs, and speculations) in texts (Pang and Lee, 2008; Wiebe, 2000; Pang, Lee, and Vaithyanathan, 2002). While Subjectivity Analysis deals with detecting the presence of subjective (versus objective) expressions in text, Sentiment Analysis deals with classifying such identified phenomena into different classes of polarity (usually positive, negative and neutral).

This paper present the first participation of the OPTIMA[1] team in TASS[2]. TASS is an experimental evaluation workshop for sentiment analysis and online reputation management systems developed with a focus on Spanish.

For the TASS 2013 edition, we only participated in the first task, entitled *sentiment analysis at global level*. In this task, the participants were asked to assess the global polarity of short texts extracted from Twitter by using 5 levels of sentiment (very positive, positive, neutral, negative and very negative), plus discriminate them from the objective ones. To tackle this task, we applied an approach based on machine learning by trying different feature combinations, using dictionary-based features and adding external data for training obtained through machine translation. The main motivation for the experiments in the TASS competition was to evaluate the manner in which our approach (applied for English and combinations of data from different languages) could perform for Spanish. The results obtained show that the use of supervised learning with additional dictionary features and external training data obtained from machine translated texts might be considered good strategies to generate learning data for polarity classification systems.

The rest of the paper is organized as follows: the following section deals with the state of the art in sentiment analysis. The main features of the proposed approaches are presented in Section 3. Section 4 describes the data used for learning while the different experiments carried out are expounded in Section 5. Finally, the results obtained and the conclusions are discussed in Section 6 and Section 7, respectively.

2. State of the art

Go, Bhayani, and Huang (2009) performed one of the first studies involving sentiment analysis applied to tweets. The authors introduced emoticons (e.g. ":)", ":(", etc.) as markers of positive and negative tweets. Following their initial findings, Read (2005) employed the method to generate a corpus of sentiment-annotated tweets. They considered that positive tweets were the ones containing with positive emoticons (e.g. ":)"), and negative tweets were the ones with negative emoticons (e.g. ":("). In their subsequent experiments, they introduce different supervised approaches (SVM, Naïve Bayes and Maximum Entropy) and various sets of features and conclude that the simple use of unigrams leads to good results, but it can be slightly improved by the combination of unigrams and bigrams.

In the same line of thinking, Pak and Paroubek (2010) also generated a corpus of tweets for sentiment analysis, by selecting positive and negative tweets based on the presence of specific emoticons. Subsequently, they perform different experiments to classify sentiment in the obtained corpus and conclude that the best settings include the use of a Naïve Bayes classifier with unigrams and part-of-speech tags.

Another approach on sentiment analysis in tweet is that of Zhang et al. (2011). Here, the authors adopt 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. The authors conclude that the most important features are those corresponding to sentiment-bearing words. Finally, Jiang et al. (2011) classify sentiment expressed on previously-given "targets" in tweets. They add information on

[1] Open Source Text Information Mining and Analysis (OPTIMA), `http://ipsc.jrc.ec.europa.eu/?id=179`

[2] Workshop on Sentiment Analysis at SEPLN, `http://www.daedalus.es/TASS2013`

the context of the tweet to its text (e.g. the event that it is related to). Subsequently, they use SVM and General Inquirer and perform a three-way classification (positive, negative, neutral).

In SemEval 2013, a task was organized on sentiment analysis in tweets (Wilson et al., 2013). Here, the best-performing systems used additional dictionaries that were built from large data sets and word-emotion association dictionaries built from millions of tweets. From here, we can see that the use of dictionaries to improve the features used in supervised learning is a good strategy.

The TASS evaluation campaign has also been organized in 2012. The best participating system employed 5 classifiers to distinguish among the 5 classes of polarity (among themselves) and the objective class (separately).

3. Proposed approaches

Two main approaches were proposed for the experiments carried out in TASS. For the ***dictionary-based*** approach, we took into account the linguistic peculiarities of tweets, regarding spelling, use of slang, punctuation, etc., and also the sentiment-bearing words from the training data were replaced with a unique label. In this way, the sentence "I love roses." will be equivalent to the sentence "I like roses.", because "like" and "love" are both positive words according to the General Inquirer[3] dictionary. If the first sentence is contained in the training data and the second sentence is contained in the test data, replacing the sentiment-bearing word with a general label increases the chance to have the second sentence classified correctly.

In the same line of thought, we also replaced modifiers with unique corresponding labels. The sentiment dictionary generated by Steinberger et al. (2011) was used to replace the sentiment-bearing words contained in the tweets with unique labels describing their polarity. As such, words that were found in the "High positive" category in the dictionary were replaced with the label "HPOSITIVE", those that were in the "Positive" category were replaced with the label "POSITIVE" and similarly for those in the "High negative" and "Negative" categories. In the same manner, negators, intensifiers and diminishers, as identified by the sentiment dictionary, were replaced with the labels "NEGATOR", "INTENSIFIER" and "DIMINISHER". Finally, we replaced the emoticons with the sentiment value they had, giving them positive, high positive, negative or high negative labels and replaced the repeated punctuation signs "!", "?", "." (more than 2 appearances) with the unique labels "MULTIEXCLAMATION", "MULTIINTERROGATION" and "MULTISTOP". As can be seen from the results obtained, although the dictionaries used (for Spanish) has many less entries than the ones (for English) originally employed in our SemEval participation (Balahur, 2013), this approach was the best-performing one.

For the second approach we also tested the performance of the sentiment classification by applying ***4 separate pairs of classifiers*** (for objective versus subjective, positive versus negative versus neutral, positive versus very positive and negative versus very negative). Our aim was to see if the use of separate classifiers (in a cascade) could improve the performance, by fitting the data more appropriately.

Finally it is noteworthy that for both approaches we used simple heuristics to select the features. Although feature selection algorithms are easy to apply when a data mining environment is used, the final choice is influenced by the data at hand and it is difficult to apply on new sets of data. After performing various tests, we chose to select the features to be employed in the classification model based on the condition that they should occur at least twice in the training set.

4. Data

Several data sets have been used to carry out the experiments. They are briefly presented in the next subsection. Further on, different pre-processing steps applied to these data sets are also explained.

4.1. Data sets

Two main data sets for learning purposes have been used in our experiments: the *general corpus* training set of TASS 2013 and the dataset of tweets used in Task 2 (B) of the SemEval 2013 evaluation campaign. The first one was provided by the TASS 2013 organizers for the *sentiment analysis at glo-*

[3] `http://www.wjh.harvard.edu/~inquirer`

bal level task[4]. This corpus contains 7,219 Twitter messages written in Spanish about well-known personalities in politics, economics, communication or culture, between November 2011 and March 2012. Each message is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. Five levels have been defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE). In addition, there is also an indication of the level of agreement or disagreement of the expressed sentiment within the content, with two possible values: AGREEMENT and DISAGREEMENT. This is especially useful to make out whether a neutral sentiment comes from neutral keywords or else the text contains positive and negative sentiments at the same time. On the other hand, a set of topics has been defined based on the thematic areas covered by the corpus. Some examples are *politics*, *soccer*, *literature* or *entertainment*. Each message has been assigned to one or several of these topics.

The sentiment-annotated tweets in the SemEval 2013 Task 2(B) were provided by the task's organizers[5]. This corpus consists of about 12,000 twitter messages covering a wide range of topics, such as known entities (e.g., Gadafi, Steve Jobs), products (e.g., kindle, android phone), and events (e.g., Japan earthquake, NHL playoffs). This data set was provided in English but we obtained the Spanish translation by applying a machine translating (MT) using the Google MT engine[6]. Therefore, this translated version was included in the training data set for our TASS experiments.

4.2. Data pre-processing

The training data have been preprocessed discarding the stop words and by applying a stemming process. For removing the stop words, the list for the Spanish language provided by Snowball[7] has been used. This list was revised manually, by discarding some words that might have influence in the polarity, such as *no*, *sí*, *más*, *mucho*... In total, from the 325 stop words included in the original Snowball list, 228 were removed, remaining a final list of 97 stop words.

Regarding the stemming process, the 3.2 version of TreeTagger[8] for Spanish has been used. This tool allows to unify different variants of the same type of word by using one unique lemma even for different gender. Some examples of these conversions are: *salgo =>salir*, *ayudar me/nos =>ayudar yo nosotros*, *pensando =>pensar*, *la que se va =>el que se ir*, *lo intento =>el intento*, *buen día =>bueno día*, etc.

Taking into account these pre-processing steps, several training data sets were generated:

- tassTrain-base: original training TASS data without applying any pre-processing step
- tassTrain-lemma: original training TASS data without removing stop words but applying the stemming process
- tassTrain-lemmaStop: original training TASS data + stopper + stemmer
- semevaltassTrain-base: tassTrain-base + tweets of SemEval 2013 Task 2 (B), without applying any pre-processing step

5. *Experiments*

Different experiments have been carried out in our first participation in the TASS workshop. They are mainly based on the machine learning approach, combining different results or even by using external semantic resources like dictionaries. WEKA[9] 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 have been applied successfully in many opinion mining tasks overcoming other machine learning techniques (O'Keefe and Koprinska, 2009; Esuli and Sebastiani, 2005).

Taking into account the learning data sets explained in the previous section, three main experiments were proposed:

[4] `http://www.daedalus.es/TASS2013/corpus.php`

[5] `http://www.cs.york.ac.uk/semeval-2013/task2`

[6] `http://translate.google.com`

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

[8] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`

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

Experiment Id.	Approach
JRC-tassTrain-base-SVM	original training TASS data without applying any pre-processing step and using SVM as learning algorithm
JRC-tassTrain-base-DICT	original training TASS data without applying any pre-processing step and using dictionaries as semantic resource. SVM is used as learning algorithm
JRC-tassTrain-base-4CLS	original training TASS data without applying any pre-processing step and following the approach based on 4 pair of classifiers
JRC-tassTrain-lemma-SVM	original training TASS data without removing stop words but applying the stemming process and using SVM as learning algorithm
JRC-tassTrain-lemma-DICT	original training TASS data without removing stop words but applying the stemming process and using the dictionary-based approach. SVM is used as learning algorithm
JRC-tassTrain-lemma-4CLS	original training TASS data without removing stop words but applying the stemming process and following the approach based on 4 pair of classifiers
JRC-tassTrain-lemmaStop-SVM	original training TASS data + stopper + stemmer and using SVM as learning algorithm
JRC-tassTrain-lemmaStop-DICT	original training TASS data + stopper + stemmer and using the dictionary-based approach. SVM is used as learning algorithm
JRC-tassTrain-lemmaStop-4CLS	original training TASS data + stopper + stemmer and following the approach based on 4 pair of classifiers
JRC-semevaltassTrain-base-SVM	original training TASS data + training SemEval 2013 data set without applying any pre-processing step and using SVM as learning algorithm
JRC-semevaltassTrain-base-DICT	original training TASS data + training SemEval 2013 data set without applying any pre-processing step and using the dictionary-based approach. SVM is used as learning algorithm

Table 1: JRC experiments submitted to TASS 2013

- **SVM**. For the first experiment, we simply applied SVM as learning algorithm. As mentioned before, SVM SMO was used as learning algorithm . Thus, these experiments are represented by containing the word SVM in their title or id.
- **DICT**. For the second experiment, we applied the dictionary-based approach (see Section 3). Specifically, we have used external semantic resources such as the dictionaries provided by Steinberger et al. (2011) and General Inquirer.
- **4CLS**. For the third experiment we applied the second approach in which we combined the results of 4 separate pairs of classifiers to obtain the final polarity value. These 4 classifiers combined objective versus subjective, positive versus negative versus neutral, positive versus very positive and negative versus very negative labeled tweets from the training data sets.

According to these approaches and the different learning data sets generated, a total of 18 experiments were submitted to the TASS 2013 workshop. Table 1 summarizes what each experiment represents. For each experiment and tweet, the global polarity using 5-levels (P+, P, NEU, N, N+) was obtained. Then, the 3-level version of each experiment was obtained by considering as *P* and *N* those tweets classified as P+ and N+, respectively. The rest remained with identical labels than those obtained for the 5-level experiments.

6. *Results and discussion*

The official results for the 18 experiments submitted to TASS 2013 are shown in Table 2 and Table 3. Table 2 shows the results for the 5-level way and Table 3 for the 3-level way. The typical measures in classification tasks, such as *precision* (P), *recall* (R) and F1 have been applied to obtain these results. It is noteworthy that the official evaluation results considered the successes and failures globally, i.e., without taking into account each class, and therefore averaging P, R and F1 in total.

As we can see from the results, our approach (that was initially tailored for English data) performed well. Although the calculation of the systems' performance as accuracy is debatable, given that the classes evaluated were highly unbalanced, we can conclude that our system is robust and the performance is relatively stable. Further analysis on the per-class performance is required in order to establish which of the classes were less well distinguishable, leading to improved features in our system. Surprisingly, the 4-classifiers approach performed the lowest. In this sen-

Experiment Id.	P	R	F1
JRC-tassTrain-base-DICT	0.519	0.519	**0.519**
JRC-tassTrain-lemmaStop-SVM	0.515	0.515	0.515
JRC-tassTrain-lemmaStop-DICT	0.507	0.507	0.507
JRC-tassTrain-base-SVM	0.505	0.505	0.505
JRC-tassTrain-lemma-SVM	0.504	0.504	0.504
JRC-tassTrain-lemma-DICT	0.497	0.497	0.497
JRC-tassTrain-lemmaStop-4CLS	0.481	0.481	0.481
JRC-tassTrain-base-4CLS	0.477	0.477	0.477
JRC-tassTrain-lemma-4CLS	0.477	0.477	0.477

Table 2: JRC 5-way official results obtained in TASS 2013

Experiment Id.	P	R	F1
JRC-tassTrain-base-DICT	0.612	0.612	**0.612**
JRC-tassTrain-lemmaStop-SVM	0.608	0.608	0.608
JRC-tassTrain-lemmaStop-DICT	0.607	0.607	0.607
JRC-tassTrain-lemma-DICT	0.599	0.599	0.599
JRC-tassTrain-lemma-SVM	0.599	0.599	0.599
JRC-tassTrain-base-SVM	0.597	0.597	0.597
JRC-semevaltassTrain-base-DICT	0.590	0.590	0.590
JRC-semevaltassTrain-base-SVM	0.585	0.585	0.585
JRC-tassTrain-lemmaStop-4CLS	0.582	0.582	0.582

Table 3: JRC 3-way official results obtained in TASS 2013

se, further analysis must be done to determine at what step of the cascade the misclassification of the examples has led to a loss in accuracy.

7. Conclusions and further work

In this article, we presented our participation to the TASS 2013 evaluation campaign. We participated with a system that was designed for English and adapted it to Spanish, by employing in-house built dictionaries and machine-translated data for training. Additionally, we tested the manner in which 4 separate classifiers could be used in cascade to determine the sentiment, from the general "'subjective" versus "objective", to the finergrained classes of polarity. Although this is our first participation, the results obtained were promising.

Several conclusions can be inferred from the experiments carried out. One of them concerns with the use of minimal linguistic processing, which makes the approach easily portable to other languages. On the other hand, from the results obtained, it has shown that the use of linguistic processing (e.g. lemmatization, stopword removal) actually worsen the performance. Finally, the use of unigrams and bigrams to spot modifications in the polarity of the sentiment expressed, allowed us to learn general patterns of sentiment expression (e.g. "negation positive", "intensifier negative", etc.). This pattern was successfully applied for English and, as we could see from the results obtained, also for Spanish.

In further work, we would like to tune our classifiers for the Spanish data employed and use additional language-specific features. We also plan to enrich the sentiment dictionaries used for the TASS experiments, so that more informal sentiment expressions can be captured and adopted as features for the polarity classification. Finally, we plan to include text normalization techniques adapted to Spanish. This was previously achieved for English (in our participation to SemEval), but due to time restrictions could not be achieved for TASS.

References

Balahur, A. 2013. OPTWIMA: Comparing knowledge-rich and knowledge-poor approaches for sentiment analysis in short informal texts. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

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.

Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Pak, Alexander and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta; ELRA, may. European Language Resources Association. 19-21.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

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*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steinberger, Josef, Polina Lenkova, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Vanni Zavarella, and Silvia Vázquez. 2011. Creating sentiment dictionaries via triangulation. In *Proceedings of WASSA 2011*, WASSA '11, pages 28–36. ACL.

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

Wiebe, Janyce. 2000. Learning subjective adjectives from corpora. In *Proceedings of AAAI.*

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.