TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 35-39

JACERONG at TASS 2016: An Ensemble Classifier for Sentiment Analysis of Spanish Tweets at Global Level

JACERONG en TASS 2016: Combinación de clasificadores para el análisis de sentimientos de tuits en español a nivel global

Jhon Adrián Cerón-Guzmán
Santiago de Cali, Valle del Cauca, Colombia
jadrian.ceron@gmail.com

Resumen: Este artículo describe un enfoque basado en conjuntos de clasificadores que se ha desarrollado para participar en la Tarea 1 del taller TASS sobre análisis de sentimientos de tuits en español a nivel global. Los conjuntos se construyen sobre la combinación de sistemas con la correlación absoluta más baja entre sí. Estos sistemas son capaces de tratar con formas léxicas no estándar en los tweets, con el fin de mejorar la calidad del análisis de lenguaje natural. Para realizar la clasificación de polaridad, el enfoque utiliza características básicas que han probado su poder discriminativo, así como características de n-gramas de palabras y caracteres. Luego, las salidas de clasificadores de Regresión logística, que pueden ser etiquetas de clase o probabilidades para cada clase, se utilizan para construir conjuntos de clasificadores. Los resultados experimentales muestran que la combinación menos correlacionada de 25 sistemas, la cual elige la clase con la probabilidad promedio no poderada más alta, es la configuración que mejor se adapta a la tarea, alcanzando una precisión global de 62.0% en la evaluación de seis etiquetas, y de 70.5% en la evaluación de cuatro etiquetas.
Palabras clave: Análisis de sentimientos, clasificación de polaridad, combinación de clasificadores, normalización léxica, tuis en español, Twitter

Abstract: This paper describes an ensemble-based approach developed to participate in TASS-2016 Task 1 on sentiment analysis of Spanish tweets at global level. Ensembles are built on the combination of systems with the lowest absolute correlation with each other. The systems are able to deal with non-standard lexical forms in tweets, in order to improve the quality of natural language analysis. To support the polarity classification, the approach uses basic features that have proved their discriminative power, as well as word and character n-gram features. Then, outputs from Logistic Regression classifiers, which may be either class labels or probabilities for each class, are used to build ensembles. Experimental results show that the less-correlated combination of 25 systems, which chooses the class with the highest unweighted average probability, is the setting that best suits to the task, achieving an overall accuracy of 62.0% in the six-labels evaluation, and of 70.5% in the four-labels evaluation.
Keywords: Ensemble classifier, lexical normalization, polarity classification, sentiment analysis, Spanish tweets, Twitter

1 Introduction

What people say on social media about issues of their everyday life, the society, and the world in general, has turned into a rich source of information to understand social behavior. Twitter content, in particular, has caught the attention of researchers who have investigated its potential for conducting studies on the human subjectivity at large scale, which was not feasible using traditional methods. Around election time, sentiment analysis of political tweets has been widely used to capture trends in public opinion regarding important issues such as voting intention (Gayo-Avello, 2013). However, analyzing this content also presents several challenges, including the development of text analysis approaches based on Natural Language Processing techniques, which properly adapt to the informal genre and the free writ-

ISSN 1613-0073

ing style of Twitter (Han and Baldwin, 2011; Cerón-Guzmán and León-Guzmán, 2016).

TASS is a workshop aimed at fostering research on sentiment analysis of Spanish Twitter data, which provides a benchmark evaluation to compare the latest advances in the field (García-Cumbreras et al., 2016). One of the proposed tasks is to determine the opinion orientation expressed in tweets at global level. Task 1 consists on assigning one of six labels (P+, P, NEU, N, N+, NONE) to a tweet in the six-labels evaluation; or one of four labels (P, NEU, N, NONE) in the four-labels evaluation. Here, P, N, and NEU, stand for positive, negative, and neutral, respectively; NONE, instead, means no sentiment. The "+" symbol is used as intensifier.

This paper presents an ensemble-based approach to polarity classification of Spanish tweets, developed to participate in Task 1 proposed by the organizing committee of the TASS workshop. The ensemble members are (relatively) highly correct classifiers with the lowest absolute correlation with each other. The output from each classifier, which may be either a class label or probabilities for each class, is used to assign the polarity to a tweet based on a majority rule or on the highest unweighted average probability. Moreover, classifiers are adapted to deal with non-standard lexical forms in tweets, in order to improve the quality of natural language analysis.

The remainder of this paper is organized as follows. Section 2 describes the common architecture of the ensemble members (i.e., classifiers). Next, the submitted experiments, as well as the obtained results, are discussed in Section 3. Finally, Section 4 concludes the paper.

2 The System Architecture

The tweet text is passed through the pipeline of each system in order to assign it a class label or a probability to be of a certain class. The pipeline, which goes from text preprocessing to machine learning classification, is described below. Note that the system term is preferred over the classifier term, because a machine learning classifier receives a feature vector and produces a class label or probabilities for each class; instead, the system term enables to conceive the whole process, from preprocessing to machine learning classification.

2.1 Preprocessing

The process of text cleaning and normalization is performed in two phases: basic preprocessing and advanced preprocessing.

2.1.1 Basic Preprocessing

The following simple rules are implemented as regular expressions:

- Removing URLs and emails.
- HTML entities are mapped to textual representations (e.g., "&lt;" → "<").
- Specific Twitter terms such as mentions (@user) and hashtags (#topic) are replaced by placeholders.
- Unknown characters are mapped to their closest ASCII variant, using the Python *Unidecode* module for the mapping.
- Consecutive repetitions of a same character are reduced to one occurrence.
- Emoticons are recognized and then classified into positive and negative, according to the sentiment they convey (e.g., ":)" → "EMO_POS", ":(" → "EMO_NEG").
- Unification of punctuation marks (Vilares, Alonso, and Gómez-Rodrıguez, 2014).

2.1.2 Advanced Preprocessing

Once the set of simple rules has been applied, the tweet text is tokenized and morphologically analyzed by FreeLing (Padró and Stanilovsky, 2012). In this way, for each resulting token, its lemma and Part-of-Speech (POS) tag are assigned. Taking these data as input, the following advanced preprocessing is applied:

- **Lexical normalization.** Each token is passed through a set of basic modules of FreeLing (e.g., dictionary lookup, suffixes check, detection of numbers and dates, and named entity recognition) for identifying standard word forms and other valid constructions. If a token is not recognized by any of the modules, it is marked as out-of-vocabulary (OOV) word. Then, a confusion set is formed by normalization candidates which are identical or similar to the graphemes or phonemes that make the

OOV word. These candidates are elements of the union of a dictionary of Spanish standard word forms and a gazetteer of proper nouns. The best normalization candidate for the OOV word is which best fits a statistical language model. The language model was estimated from the Spanish Wikipedia corpus. Lastly, the selected candidate is capitalized according to the capitalization rules of the Spanish language. Extensive research on lexical normalization of Spanish tweets can be read in (Cerón-Guzmán and León-Guzmán, 2016).

- **Negation handling.** Inspired by the approach proposed by Pang et al. (Pang, Lee, and Vaithyanathan, 2002), this research defined a negated context as a segment of the tweet that starts with a (Spanish) negation word and ends with a punctuation mark (i.e., "!", ",", ":", "?", ".", ";"), but only the first $n \in [0,3]$ or all tokens labeled with any or a specific POS tag (i.e., verb, adjective, adverb, and common noun) are affected by adding it the "_NEG" suffix. Note that when $n = 0$, no token is affected.

2.2 Feature Extraction

In this stage, the normalized tweet text is transformed into a feature vector that feeds the machine learning classifier. The features are grouped into basic features and n-gram features.

2.2.1 Basic Features

Some of these features are computed before the process of text cleaning and normalization is performed.

- The number of words completely in uppercase.
- The number of words with more than two consecutive repetitions of a same character.
- The number of consecutive repetitions of exclamation marks, question marks, and both punctuation marks (e.g., "!!", "??", "?!") and whether the text ends with an exclamation or question mark.
- The number of occurrences of each class of emoticons (i.e., positive and negative) and whether the last token of the tweet is an emoticon.
- The number of positive and negative words, relative to the ElhPolar lexicon (Saralegi and Vicente, 2013), the AFINN lexicon (Nielsen, 2011), or an union of both lexicons. In a negated context, the label of a polarity word is inverted (i.e., positive words become negative words, and vice versa). Additionally, a third feature labels the tweet with the class whose number of polarity words in the text is the highest.
- The number of negated contexts.
- The number of occurrences of each Part-of-Speech tag.

2.2.2 N-gram Features

The fixed-length set of basic features is always extracted from tweets. However, the tweet text varies from another in terms of length, number of tokens, and vocabulary used. For that reason, a process that transforms textual data into numerical feature vectors of fixed length is required. This process, known as vectorization, is performed by applying the tf-idf weighting scheme (Manning, Raghavan, and Schütze, 2008). Thus, each document (i.e., a tweet text) is represented as a vector $d = \{t_1, \ldots, t_n\} \in \mathbb{R}^V$, where V is the size of the vocabulary that was built by considering word n-grams with $n \in [1,4]$, or character n-grams with $n \in [3,5]$ in the collection (i.e., the training set). The vector is, hence, formed by word n-grams, character n-grams, or a concatenation of word and character n-grams.

2.3 Machine Learning Classification

At the last stage, the sentiment analysis system classifies a given tweet as either P+, P, NEU, N, N+, or NONE, or assigns probabilities for each class. After receiving as input the feature vector, a L2-regularized Logistic Regression classifier assigns a class label to the tweet or a probability to be of a certain class. The classifier was trained on the training set, using the Scikit-learn (Pedregosa et al., 2011) implementation of the Logistic Regression algorithm.

3 Experiments

1,720 different sentiment analysis systems were trained on the training set via 5-fold cross validation, in order to find the best parameter settings, namely: negation handling, polarity lexicon, order of word and character n-grams, and others parameters related to the vectorization process (e.g., lowercasing, frequency thresholds, etc.). The systems were sorted by their mean cross-validation score, and thus the top 50 ranked were filtered to build the ensemble. The training set is a collection of 7,219 tweets, each of which is tagged with one of six labels (i.e., P+, P, NEU, N, N+, and NONE). Note that the systems were trained for the six-labels evaluation, and therefore the P+ and P labels were merged into P, as well as the N+ and N labels were merged into N, to produce an output in accordance with the four-labels evaluation. Further description of the provided corpus, as well as of the training and test sets, can be read in (García-Cumbreras et al., 2016).

Next, the top 50 systems assigned a class label to each tweet in a collection of 1,000, which was drawn from the untagged test set with a similar class distribution to the training set. In this stage, the objective was to find the systems with the lowest absolute correlation with each other; therefore, the performance was not evaluated. Then, the less-correlated combinations of 5, 10, and 25 systems, were used to build the ensembles, whose outputs correspond to the submitted experiments. These experiments are described below:

- **run-1**: the less-correlated combination of 5 systems, which chooses the class label that represents the majority in the predictions made by the ensemble members.
- **run-2**: the less-correlated combination of 10 systems, which chooses the class with the highest unweighted average probability.
- **run-3**: the less-correlated combination of 25 systems, which chooses the class with the highest unweighted average probability.

Tables 1 and 2 show the performance evaluation on the test set (i.e., a collection of 60,798 tweets) for six and four labels, respectively. Accuracy has been defined as the official metric for ranking the systems. In summary, the main gain occurs among the "run-1" and "run-2" experiments, with an increment of 0.5% in accuracy in the six-labels evaluation, and of 0.2% in the four-labels evaluation; instead, a negligible gain occurs among the "run-2" and" run-3" experiments, taking additionally into account the computational cost of running the latter.

Experiment	**Accuracy**	**Macro-Precision**	**Macro-Recall**	**Macro-F1**
run-1	0.614	0.471	0.531	0.499
run-2	0.619	0.476	0.535	0.504
run-3	0.620	0.477	0.532	0.503

Table 1: Performance on the test set in the six-labels evaluation

Experiment	**Accuracy**	**Macro-Precision**	**Macro-Recall**	**Macro-F1**
run-1	0.702	0.564	0.565	0.564
run-2	0.704	0.567	0.568	0.567
run-3	0.705	0.568	0.567	0.568

Table 2: Performance on the test set in the four-labels evaluation

Class	**Precision**	**Recall**	**F1-score**
P	0.755	0.786	0.770
NEU	0.128	0.093	0.107
N	0.631	0.812	0.710
NONE	0.758	0.578	0.656

Table 3: Discriminative power for each class in the four-labels evaluation

As a final point, Table 3 shows how the overall performance is affected by the low discriminative power of the ensembles (in this case, the one that correspond to "run-3") for the NEU class. With this in mind, it is proposed as future work to deal with the low representativeness of the NEU class in the training data (i.e., 9.28% of tweets), in order to properly characterize this kind of tweets.

4 Conclusion

This paper has described an ensemble-based approach for sentiment analysis of Spanish Twitter data at global level, developed in order to participate in Task 1 proposed by the organization of TASS workshop. Three ensembles were built on the combination of sentiment analysis systems with the lowest absolute correlation with each other. The systems were adapted to the informal genre and the free writing style that characterize Twitter, in order to improve the quality of natural language analysis. In this way, the predicted class label for a particular tweet was based on a majority rule or on the highest average probability. Experimental results showed that the less-correlated combination of 25 systems, which chose the class with the highest unweighted average probability, was the setting that best suited to the task. However, there is a great room for improvement in the learning of a proper characterization of neutral tweets.

References

Cerón-Guzmán, J. A. and E. León-Guzmán. 2016. Lexical normalization of Spanish tweets. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW'16 Companion, pages 605–610. International World Wide Web Conferences Steering Committee.

García-Cumbreras, M. A., J. Villena-Román, E. Martínez-Cámara, M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Urena-López. 2016. Overview of tass 2016. In *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016)*, Salamanca, Spain, September.

Gayo-Avello, D. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Soc. Sci. Comput. Rev.*, 31(6):649–679.

Han, B. and T. Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT'11, pages 368–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. Scoring, term weighting and the vector space model. In *An Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Nielsen, F. Å. 2011. A new anew: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86. Association for Computational Linguistics.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Saralegi, X. and I. S. Vicente. 2013. Elhuyar at tass 2013. In *Proceedings of the Sentiment Analysis Workshop at SEPLN (TASS2013)*, September.

Vilares, D., M. A. Alonso, and C. Gómez-Rodrıguez. 2014. On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *Journal of the Association for Information Science and Technology.*