TASS 2017: Workshop on Semantic Analysis at SEPLN, septiembre 2017, págs. 59-64

Classifier Ensembles That Push the State-of-the-Art in Sentiment Analysis of Spanish Tweets

Conjuntos de clasificadores que mejoran el estado del arte para el análisis de sentimientos de tuits en español

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

Abstract: This paper describes the JACERONG system proposed to participate in TASS-2017 Task 1. For such a benchmark evaluation, two ensemble methods widely utilized because of their proved ability to increase prediction accuracy were implemented, namely: averaging and stacking. First of all, (relatively) highly correct classifiers utilize supervised learning algorithms to predict a class label or probability estimates. Then, predictions from these classifiers are optimally combined in order to obtain a better final prediction. Finally, how to choose which classifiers constitute an ensemble was also an important issue addressed in this work. Experimental results show that the proposed system is top-ranked on the test set of the InterTASS corpus, according to the accuracy metric. Together with this, results indicate that the predictive performance on the whole test set of the General Corpus of TASS outperforms the best result achieved in the four-label evaluation of the previous edition of TASS, in terms of the Macro-F1 metric.
Keywords: Classifier ensembles, sentiment analysis, Spanish tweets, Twitter

Resumen: Este artículo describe el sistema JACERONG propuesto para participar en la Tarea 1 de TASS 2017. Para tal evaluación, se implementaron dos métodos de combinación de clasificadores ampliamente utilizados debido a su demostrada capacidad de aumentar la exactitud de predicción, a saber: promediar y apilar. En primer lugar, clasificadores (relativamente) muy correctos utilizan algoritmos de aprendizaje supervisado para predecir una etiqueta de clase o estimaciones de probabilidad. Luego, se combinan de manera óptima las predicciones de estos clasificadores con el fin de obtener una mejor predicción final. Por último, también se exploró cómo elegir cuáles clasificadores constituyen un conjunto. Los resultados experimentales muestran que el sistema propuesto es el mejor clasificado en la evaluación del corpus InterTASS, de acuerdo con la métrica oficial de exactitud. Asimismo, los resultados indican que el desempeño predictivo sobre el conjunto de evaluación completo del Corpus General de TASS es superior al mejor resultado alcanzado en la evaluación de cuatro etiquetas de la edición anterior de TASS, en términos de la métrica oficial Macro-F1.
Palabras clave: Análisis de sentimientos, conjuntos de clasificadores, tuits en español, Twitter

1 Introduction

Nowadays, 'tweeting' has become an activity *par excellence* to say what one thinks or feels. Thus, the large amount of user-generated content on Twitter, in the form of short texts limited to 140 characters that are known as tweets, has led to develop new methods to explore the human subjectivity at large scale. Sentiment analysis, as one of these methods is known, has been widely utilized to gauge public opinion regarding important issues of people's everyday life, the society, and the word in general, e.g. a political election (Cerón-Guzmán and León-Guzmán, 2016b); it also benefits from the exponential growth of the computational capacity to process such a large volume of information.

TASS is a workshop aimed at fostering research on sentiment analysis of Spanish tweets, which provides a benchmark evalu-

ISSN 1613-0073

ation to compare the latest advances in the field (Martínez-Cámara et al., 2017). One of the proposed tasks is to determine the opinion orientation expressed at tweet level. Task 1 consists in assigning one of four labels (P, NEU, N, NONE) to a given tweet. Here, P, N, and NEU, stand for positive, negative, and neutral, respectively; NONE, instead, means no sentiment.

This paper describes the JACERONG system proposed to participate in TASS-2017 Task 1. For this sixth edition, classifier ensembles based on stacking were developed, in addition to the ones based on averaging, with several improvements, that were presented in the previous edition (Cerón-Guzmán, 2016). Regarding the ensembles, they are constituted by (relatively) highly correct classifiers that utilize Logistic Regression and Support Vector Machines as the supervised learning algorithms to predict a class label or probability estimates. Then, predictions from these classifiers are optimally combined in order to obtain a better final prediction. Finally, how to choose which classifiers constitute an ensemble was also an important issue addressed in this work.

The remainder of this paper is organized as follows. Section 2 explains the system architecture. Next, the submitted runs and the obtained results are discussed in Section 3. Lastly, Section 4 concludes the paper.

2 *The System Architecture*

The system architecture can be viewed as a pipeline consisting of several pre-processing modules, a vectorizer that transforms a text into a feature vector, machine learning classifiers, and an ensemble combiner that takes level-one predictions and then optimally combines them to obtain a better final prediction. Figure 1 illustrates the system architecture. In addition to this, code of the system is publicly available to enable reproducibility.[1]

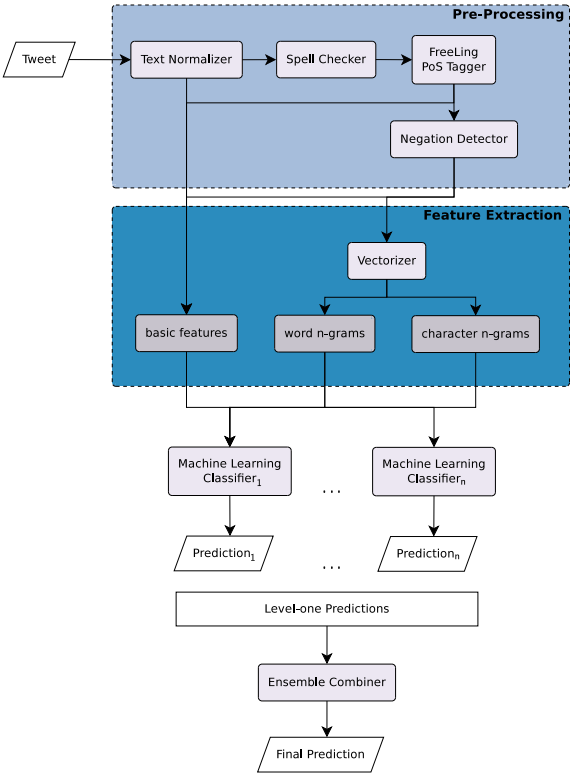

Figure 1: The system architecture

2.1 Pre-processing

2.1.1 Text Normalizer

This is a rule-based normalizer as listed below:

- Removing URLs and emails.
- HTML entities are mapped to their textual representation (e.g., "<" → "<").
- 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, 2015).

2.1.2 Spell Checker

An open-source spell checker for Spanish texts is used to normalize non-standard word forms, i.e. out-of-vocabulary (OOV) words, to their standard lexical form (Cerón-Guzmán and León-Guzmán, 2016a).[2] Normalesp suggests normalization candidates that are identical or similar to the graphemes or phonemes that make an OOV word, and

[1] https://github.com/jacerong/TASS-2017

[2] https://github.com/jacerong/normalesp

using contextual information, it selects the best normalization candidate.

2.1.3 Negation Detector

Inspired by the approach proposed by Pang et al. (Pang, Lee, and Vaithyanathan, 2002), a negated context is defined as a segment of the text that starts with a negation word and ends with a punctuation mark (i.e., "!", ",", ":", "?", ".", ";"), but only the first $n \epsilon [0, 3]$ 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. The negation detector uses FreeLing (Padró and Stanilovsky, 2012) to tokenize the text and assign Part-of-Speech (POS) tags to the resulting tokens.

2.2 Feature Extraction

Once the text has been normalized as described above, it is transformed into a feature vector that feeds a first-level classifier. The feature vector is formed by concatenating basic and n-gram features.

2.2.1 Basic Features

Some of the following features are computed before the text 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 text is an emoticon.
- The number of positive and negative words, relative to the ElhPolar lexicon (Saralegi and Vicente, 2013), the AFINN lexicon (Nielsen, 2011), the iSOL lexicon (Molina-González et al., 2013), the EmoLex Lexicon (Mohammad and Turney, 2013), the StrengthLex lexicon (Pérez-Rosas, Banea, and Mihalcea, 2012), or a union of two, three, four, or all lexicons. In a negated context, the polarity of a 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 a text. However, a text varies from another in terms of length, number of tokens, and vocabulary. 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 $d = \{t_1, \ldots, t_n\} \epsilon \mathbb{R}^V$, where V is the size of the vocabulary that was built by considering word n-grams with $n \epsilon [1, 4]$, or character n-grams with $n \epsilon [2, 5]$ 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 Classifier

At this stage, a machine learning classifier, or first-level classifier, receives the feature vector and predicts a class label or probability estimates, i.e. the probability of the tweet to be of a certain class. Whichever the prediction be, it is denominated level-one prediction. Logistic Regression and Support Vector Machines (SVM) with 'linear' kernel are the algorithms utilized to develop a supervised learning classification approach; Scikit-learn (Pedregosa et al., 2011) is the machine learning library used.

2.4 Ensemble Combiner

Two ensemble methods were implemented to take level-one predictions and then optimally combine them in order to obtain a better final prediction, namely: averaging and stacking (Li et al., 2014). The former chooses the class with the highest unweighted average probability from probability estimates predicted by first-level classifiers. In spite of its simplicity, it has proved to be a competitive method that allows to achieve top results (Cerón-Guzmán,

2016). Regarding the latter, it stacks class labels predicted by first-level classifiers and then provides them as input to a second-level classifier to generate an ensemble prediction, i.e. the final prediction. SVM with 'radius basis function' kernel is the algorithm utilized to generate final predictions.

3 Experiments

Firstly, the training data were used to fit 8,774 first-level classifiers (4,387 of which were learned from the training set of the InterTASS corpus, while to learn the remaining ones the training set of the General Corpus of TASS was used) via 5-fold cross validation in order to find the best parameter settings, namely: scope of the negated context, polarity lexicon, order of word and character n-grams, and other parameters related to the vectorizer (e.g., frequency thresholds). Secondly, these classifiers were ranked according to their predictive performance on cross validation, i.e. the (out-of-fold) prediction accuracy obtained by averaging among the k iterations; out-of-fold predictions in the k-th iteration are the predictions obtained by applying a first-level classifier, which was trained on $k-1$ folds, to the remaining one subset. Thus, the best 100 first-level classifiers for each training set were filtered. Thirdly, how to choose which first-level classifiers constitute an ensemble was an important issue tackled in this work. Empirical findings indicate that the less-correlated combination of classifiers achieves top results (Cerón-Guzmán, 2016). Finally, second-level classifiers were trained using out-of-fold predictions on cross validation. Regarding this matter, only ensembles based on stacking were trained via 5-fold cross validation.

In order to evaluate the predictive performance of the system, the test set of the InterTASS corpus and the two test sets of the General Corpus of TASS (the whole set and the stratified sample of 1,000 tweets) were used. Specifically, given a tweet from any of the test sets, its polarity should be predicted; the polarity, or class label, can be P, N, NEU, or NONE. Macro-F1 and Accuracy are the official metrics used to rank the participating systems. Regarding the provided corpora, and the way these are split into training and test sets, the reader is referred to (Martínez-Cámara et al., 2017) where they are thoroughly described.

Experiment	Whole Set		1K Set	
	Macro-F1	Accuracy	Macro-F1	Accuracy
legacy-run-1	0.569 **(2)**	0.706 **(2)**	0.508	0.678 **(2)**
legacy-run-2	0.545	0.701	0.506	0.673
legacy-run-3	0.568	0.705	0.518 **(5)**	0.625

Table 1: Overall performance on the test sets of the General Corpus of TASS

Experiment	Macro-F1	Accuracy
InterTASS-run-1	0.459	0.608 **(1)**
InterTASS-run-2	0.460 **(4)**	0.602
InterTASS-run-3	0.430	0.576

Table 2: Overall performance on the test set of the InterTASS corpus

Tables 1 and 2 show the obtained results of the runs submitted to evaluate the predictive performance of the JACERONG system on the test sets of the General Corpus of TASS and the test set of the InterTASS corpus, respectively. The integers in parentheses correspond to the official ranks achieved by the proposed system in TASS-2017 Task 1, according to the best result for each metric. Concerning the runs submitted to evaluate the system on the two test sets of the General Corpus of TASS, these are described below:

- **legacy-run-1**: the less-correlated combination of 14 first-level classifiers learned from the training set of the General Corpus of TASS, which constitute an ensemble based on averaging.
- **legacy-run-2**: the less-correlated combination of 21 first-level classifiers learned from the training set of the General Corpus of TASS, which constitute an ensemble based on stacking.
- **legacy-run-3**: it is the same run submitted to TASS-2016 Task 1 that achieved the best results (Cerón-Guzmán, 2016), namely: the less-correlated combination of 25 first-level classifiers learned from the training set of the General Corpus of TASS, which constitute an ensemble based on averaging.

In the same way, the runs submitted to evaluate the system on the test set of the InterTASS corpus are described below:

- **InterTASS-run-1**: the less-correlated combination of 3 first-level classifiers

learned from the training set of the InterTASS corpus, which constitute an ensemble based on averaging.

- **InterTASS-run-2**: the less-correlated combination of 19 first-level classifiers learned from the training set of the InterTASS corpus, which constitute an ensemble based on stacking.

- **InterTASS-run-3**: the less-correlated combination of 14 first-level classifiers learned from the training set of the General Corpus of TASS, which constitute an ensemble based on averaging.

In summary, it is worth to state that ensembles based on averaging are significantly better than the ones based on stacking. And this significance does not only correspond to the slightly better results achieved by the former, but also to their ability to increase prediction accuracy given their simplicity and their computational efficiency. Thus, the proposed system outperforms all the participating systems in predictive performance on the test set of the InterTASS corpus, in terms of the accuracy metric; likewise, the predictive performance on the whole test set of the General Corpus of TASS turns out to be slightly better than the best result achieved in the four-label evaluation of the previous edition (García-Cumbreras et al., 2016), in terms of the Macro-F1 metric. Additionally, the obtained results of the third run submitted to evaluate the system on the InterTASS corpus ("InterTASS-run-3") should be highlighted, taking into account that the domain from which the first-level classifiers that constitute the ensemble were learned differs from the one of evaluation; specifically, these results are above-average (0.5642 in terms of the accuracy metric, taking only the best result from each participating system).

As a final point, class imbalance is a major problem that has not been properly tackled yet. Specifically, the overall performance of the system was significantly affected by the low discriminative power for the NEU class, on both the test set of the InterTASS Corpus and the two test sets of the General Corpus of TASS. With this in mind, future research efforts should be focused on dealing with the low representativeness of the NEU class.

4 Conclusion

This paper has described the JACERONG system proposed to participate in TASS-2017 Task 1. For such a benchmark evaluation, two ensemble methods were implemented, namely: averaging and stacking. Findings indicate that ensembles based on averaging are significantly better than the ones based on stacking. This significance corresponds to the former's ability to increase prediction accuracy given their simplicity and their computational efficiency, in addition to the slightly better results achieved by them. Moreover, findings show that the less-correlated combination of classifiers achieves top results. The experimental evaluation on the test set of the InterTASS corpus showed that the proposed system is top-ranked. Together with this, results indicated that the predictive performance on the whole test set of the General Corpus of TASS outperforms the best result achieved in the four-label evaluation of the previous edition of TASS.

References

Cerón-Guzmán, J. A. 2016. Jacerong at TASS 2016: An ensemble classifier for sentiment analysis of Spanish tweets at global level. In *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016)*, pages 35–39.

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

Cerón-Guzmán, J. A. and E. León-Guzmán. 2016b. A sentiment analysis system of Spanish tweets and its application in Colombia 2014 presidential election. In *Proceedings of the 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, pages 250–257.

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), pages 13–21.

Li, Y., J. Gao, Q. Li, and W. Fan. 2014. Ensemble learning. In *Data Classification: Algorithms and Applications*. Chapman & Hall/CRC.

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.

Martínez-Cámara, E., M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, and J. Villena-Román. 2017. Overview of TASS 2017. In *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Mohammad, S. M. and P. D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

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.

Pérez-Rosas, V., C. Banea, and R. Mihalcea. 2012. Learning sentiment lexicons in Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3077–3081.

Saralegi, X. and I. S. Vicente. 2013. Elhuyar at TASS 2013. In *Proceedings of the Sentiment Analysis Workshop at SEPLN (TASS2013)*, pages 143–150.

Vilares, D., M. A. Alonso, and C. Gómez-Rodrıguez. 2015. On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *Journal of the Association for Information Science and Technology*, 66(9):1799–1816.