TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 57-63

RETUYT-InCo at TASS 2018: Sentiment Analysis in Spanish Variants using Neural Networks and SVM

RETUYT-InCo en TASS 2018: Análisis de Sentimiento en Variantes del Español mediante Redes Neuronales y SVM

Luis Chiruzzo **Aiala Rosá**
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
{luischir, aialar}@fing.edu.uy

Resumen: En este artículo se presentan tres enfoques para el análisis de senimiento de tweets en diferentes variantes del español, en el marco del TASS 2018. Los clasificadores se basan en máquinas de vectores de soporte (SVM), redes neuronales convolucionales (CNN) y redes neuronales recurrentes de tipo LSTM. Si bien para las diferentes variantes se encontraron diferentes clasificadores que funcionaron mejor, en todos los casos el uso de *word embeddings* fue clave para mejorar el desempeño. Además, utilizar una técnica de entrenamiento de balance alternado para la LSTM permitió mejoras significativas en la detección de tweets neutros.
Palabras clave: Análisis de Sentimiento, LSTM, Redes Neuronales, Word Embeddings

Abstract: This paper presents three approaches for classifying the sentiment of tweets for different Spanish variants in the TASS 2018 challenge. The classifiers are based on Support Vector Machines (SVM), Convolutional Neural Netowrks (CNN) and Long Short Term Memory networks (LSTM). Although different classifiers worked better for different language variants, the use of word embeddings was key for obtaining performance improvements. Also, using a mixed-balanced training method for the LSTM resulted in a significant improvement in the detection of neutral tweets.
Keywords: Sentiment Analysis, LSTM, Neural Networks, Word Embeddings

1 Introduction

Sentiment analysis is one of the most important tasks related to subjectivity analysis within Natural Language Processing. The sentiment analysis of tweets is especially interesting due to the large volume of information generated every day, the subjective nature of most messages, and the easy access to this material for analysis and processing. The existence of specific tasks related to this field, for several years now, shows the interest of the NLP community in working on this subject. The International Workshop on Semantic Evaluation (SemEval) includes a task on Tweets Sentiment Analysis since 2013 [1]. For Spanish, the TASS workshop, organized by the SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural), focuses on this task since 2012[2].

In the TASS editions prior to 2017, most of the participants presented machine learning systems based on hand crafted features. For example, in TASS 2016 (Garcıa-Cumbreras et al., 2016) best results were obtained by a system based on an ensemble of Logistic Regression classifiers including features derived from a subjective lexicon, negation processing, and n-grams (Cerón-Guzmán, 2016); and a system based on a set of SVM classifiers with morpho-syntactic information and n-grams as features (Hurtado y Plà, 2016). Other authors (Montejo-Ráez y Díaz-Galiano, 2016; Quirós, Segura-Bedmar, y Martínez, 2016) used word embeddings, reaching lower results.

In TASS 2017 (Martínez-Cámara et al., 2017) (task 1) several systems used deep learning approaches. The best results were obtained by: Hurtado, Pla, y González (2017), who experimented with different deep neural network architectures, using as in-

[1]https://www.cs.york.ac.uk/semeval-2013/task2.html
[2]http://www.sepln.org/workshops/tass/2012/

ISSN 1613-0073

put domain-specific and general-domain sets of embeddings; Cerón-Guzmán (2017), who presented an ensemble of SVM and Logistic Regression classifiers; Rosá et al. (2017), who presented an SVM classifier based on the centroid of the tweets embeddings, a deep neural network (CNN), and a combination of both; and Moctezuma et al. (2017), who combined an SVM classifier with genetic programming.

On the other hand, for the first time, SemEval 2018 (Mohammad et al., 2018) included a dataset for Spanish tweets sentiment analysis. The corpus used in task 1.4 (ordinal classification of sentiment) is annotated with 7 values, indicating different levels of positive or negative sentiment. The best results for Spanish were obtained by systems based on deep neural networks.

In this paper we describe different approaches for Spanish tweet classification presented by the RETUYT-InCo team for the TASS 2018 sentiment analysis challenge (Martínez-Cámara et al., 2018): an SVM-based classifier which uses a set of features, including word embeddings; and two deep neural network approaches: CNN and LSTM.

2 *Corpus pre-processing*

For this year's edition of the challenge, the organizers provided three sets of corpora for Spanish variants spoken in different countries: Spain (ES), Costa Rica (CR) and Peru (PE). For each of the variants, training, development and test data was provided. The training and development sets were annotated with four possible polarity categories per tweet: `P`, `N`, `NEU` or `NONE`. The test corpora had no annotations.

For some of our experiments, we also used the general TASS training data from a previous edition of the competition. This corpus was divided in training (85 %) and development (15 %) subsets. Table 1 shows the sizes of the different corpora and the number of tweets for each class.

Each corpus was pre-processed as follows:

- Redundant space characters and ellipsis were removed.
- Twitter user references were replaced by the token "`@user`".
- URL references were replaced by the token "`@url`".

Corpus	Category	Train	Dev
General	N	1877	305
	NEU	588	82
	NONE	1207	276
	P	2464	420
	Total	6136	1083
InterTASS-ES	N	418	219
	NEU	133	69
	NONE	139	62
	P	318	156
	Total	1008	506
InterTASS-CR	N	311	110
	NEU	94	39
	NONE	165	58
	P	230	93
	Total	800	300
InterTASS-PE	N	242	106
	NEU	166	61
	NONE	361	238
	P	231	95
	Total	1000	500

Table 1: Size and categories distribution for the different corpora

- Sequences of three or more occurrences of the same character were replaced by a unique occurrence of that character. For instance, *"holaaaa"* was replaced by *"hola"*.
- Interjections denoting laughter (*"jajajaja"*, *"jejeje"*, *"jajaj"*) were replaced by the token *"jaja"*.
- The text was converted to lowercase.

We did not include any grammatical information, like lemma, POS-tag, morphological or syntactic information.

3 *Resources*

3.1 Positive and Negative Lexicons

We built a subjective lexicon consisting of the union of three subjective lexicons available for Spanish (Cruz et al., 2014; Saralegi y San Vicente, 2013; Brooke, Tofiloski, y Taboada, 2009). The lexicon, containing 6875 negative lemmas and 4853 positive lemmas, was expanded with the inflectional forms of each lemma, reaching a total of 76291 words (48959 negative and 27332 positive). This was done in order to alleviate the fact that tweets were not lemmatized. For the lexicon expansion we used the FreeLing dictionary (Padró y Stanilovsky, 2012).

In a previous work (Rosá et al., 2017), we used the same three Spanish lexicons, but we took the intersection instead of the union, obtaining a lexicon with 4730 words. Some experiments showed that the largest lexicon provides a small improvement on results when used to calculate some SVM features (as described below).

3.2 Word embeddings set

We used a 300 dimension word embeddings set, trained by (Azzinnari y Martínez, 2016) using *word2vec* (Mikolov et al., 2013). These embeddings are based on a corpus of almost six billion words in Spanish. Most of the texts come from Internet media sites.

3.3 Word Polarity Predictor

We built a regression algorithm based on SVM using the subjective lexicon as training set. This model should be able to predict a real number representing the polarity of each word. The model takes as input the 300 real values of the vector representing the word and returns a real value for the word polarity. For training, we assigned the value 1 to positive words and the value -1 to negative words. In table 2 we show the result of applying this classifier to some words.

Word	Prediction
apoyamos	1.09973945
amigo	0.89985318
excelente	1.04574863
cansancio	-0.98582263
abatían	-1.02370082
horrible	-0.91882273
apartamento	-0.30991363
teléfono	-0.48884958

Table 2: Examples of Word Polarity Prediction

As these examples show, words expected to be positive have values close to 1 and words expected to be negative have values close to -1. On the other hand, neutral words have values closer to 0 than to 1 or -1.

3.4 Category Markers

We obtained the list of all the words in the training corpus and for each one we calculated the distribution of the four categories in all the tweets where this word occurs. We consider that a word is a *category marker* if it occurs at least 75 % times in this category. Using this information we built markers lists for the four categories: 429 positive words, 438 negative words, 12 neutral words, and 33 no opinion markers.

4 Classifiers

4.1 SVM based approach

The SVM classifier configurations are almost the same as the ones described in (Rosá et al., 2017). However, the use of the new positive and negative word lexicons implied retraining the word polarity predictor and rebuilding the feature sets. These are the features used by the SVM classifiers:

- Centroid of tweet word embeddings. Previous works showed that, while using the centroid (or mean vector) is a simple technique, it reaches good results for several NLP problems, particularly for sentiment analysis (White et al., 2015). (300 real values)
- Polarity of the nine more relevant words of the tweet according to the polarity predictor. The number nine is the average length of tweets in the training corpus, filtering stop words. We considered that the more relevant words are those words whose polarities have the highest absolute value. If the tweet has less than nine words we completed the nine values repeating the polarities of the words in the tweet. (9 real values)
- Number of words belonging to the positive lexicon and to the negative lexicon. (2 natural values)
- Number of words whose vector representations are close to the mean vector of the positive and the negative lexicons. (2 natural values)
- Number of words belonging to the lists of category markers. (4 natural values)
- Features indicating if the original tweet has repeated characters or some word written entirely in upper case. (2 boolean values)
- Tentative polarity (P, N, NEU, NONE) of the tweet, based on the number of positive and negative words in the tweet,

taking into account negation markers (from a list). We inverted the polarity of words occurring between the negation marker and a punctuation mark. (4 classes)

- The five more relevant words from the training corpus, according to a bag of words classifier. The value five was experimentally defined. We filtered out words belonging to a list of stop words adapted for this task (some words relevant for sentiment analysis, such as *"no"* and *"pero"* were removed from the stop words list). (5 boolean values)

As in the previous editions, the SVM experiments were done using the *scikit-learn* toolkit (Pedregosa et al., 2011) and trained using the multiclass probability estimation method based on (Wu, Lin, y Weng, 2004).

4.2 CNN based approach

Our CNN approach uses a simpler network than the one used in (Rosá et al., 2017). In that case it was a convolutional network with three branches considering two, three and four words of context, but in our case only one convolutional branch considering three words of context was used, as shown in figure 1. The input of the network is the sequence of word embeddings corresponding to each word in the tweet, up to a maximum of 32 words. This input is fed to the convolutional layer, then the output goes to a max pooling layer and a dense layer with a dropout of 0.2 before going to a softmax layer for output. For training this network we keep a 70 %-30 % split for validation and use early stopping over the validation set.

4.3 LSTM based approach

Our LSTM neural network architecture uses the embedding for each word as input, up to a maximum of 32 words. This input is sent through a LSTM layer and then a dense layer with a dropout of 0.2, before getting the output through a softmax layer, as shown in figure 2.

The initial experiments using this network yielded good accuracy results, but the macro-F measure was very low because the network did not predict any output for the class `NEU`. This class has proven to be the most difficult to learn throughout our experiments. However, we started to get better results using

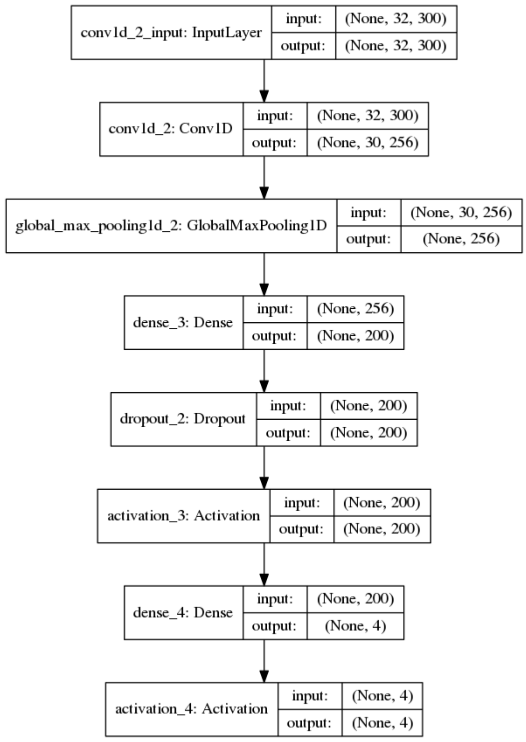

Figure 1: CNN network architecture.

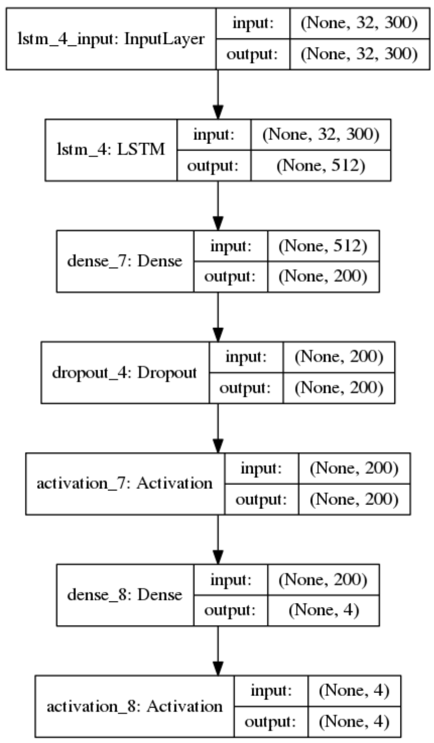

Figure 2: LSTM network architecture.

a different training strategy: we created two versions of the training corpus, one of them with all the tweets, and the other one taking the same number of tweets for each category (exactly the same number of tweets as the

NEU category, which was the one with the fewest tweets). We call this set the *balanced* corpus.

The training strategy involves training one epoch with the whole corpus and one epoch with the balanced corpus, then iterate this training process until the performance over the development set stopped improving. Training the network in this fashion yields a little less accuracy but it compensates in macro-F measure, as it captures a lot more tweets of the NEU category.

Both neural network approaches (CNN and LSTM) were implemented using the *Keras* library (Chollet, 2015) and trained using the *adam* optimization algorithm (Kingma y Ba, 2014).

5 Results

Three different corpora considering three Spanish variants were used for this task: from Spain (ES), Costa Rica (CR) and Peru (PE). Furthermore, the systems could be trained with training data for the corresponding Spanish variant (monolingual case), or they could be trained using data from other variants (cross-lingual case). We decided to submit the two best results for each classifier family on each of the variants and training combinations. Our results are shown in tables 3 and 4.

Taking in consideration the macro-F measure, our systems achieved good performance in all the test variants, ranking top 1 for monolingual CR and PE and cross-lingual ES and CR; and ranking top 2 for monolingual ES and cross-lingual PE. The best results for our systems in the monolingual training case were achieved by the neural networks approaches: in two cases, the best systems were LSTMs and in the other case it was a CNN. In the cross-lingual training cases, on the other hand, the three best systems were SVMs.

We submitted another system that combined the output probabilities of the best LSTM and SVM, in order to leverage the information of both classifiers. This approach had yielded good results in the past (Rosá et al., 2017). In this case, although the performance of the combined approach was good (49.1 % macro-F for the ES corpus), it was still a little lower than the LSTM approaches.

As can be seen in table 5, one of the reasons the LSTM could have gotten better results over the test set was because it could

		Dev		Test	
Submission	Var	Acc	F1	Acc	F1
svm_es_1	ES	57.5	45.1	**59.0**	45.9
svm_es_2	ES	57.7	45.9	58.4	47.3
cnn_es_1	ES	60.2	46.9	59.2	45.8
cnn_es_2	ES	58.1	47.1	57.4	44.5
lstm_es_1	ES	53.6	48.6	54.9	**49.9**
lstm_es_2	ES	52.4	48.9	51.4	49.8
svm_cr_1	CR	55.0	47.1	56.7	49.3
svm_cr_2	CR	58.0	46.8	**57.7**	49.9
cnn_cr_1	CR	59.3	49.6	56.9	47.7
cnn_cr_2	CR	59.3	49.6	56.3	46.9
lstm_cr_1	CR	52.3	49.6	53.0	47.3
lstm_cr_2	CR	50.7	47.2	53.7	**50.4**
svm_pe_1	PE	44.6	40.9	47.4	43.7
svm_pe_2	PE	46.8	40.2	47.1	44.1
cnn_pe_1	PE	48.2	41.8	**49.4**	**47.2**
cnn_pe_2	PE	44.0	38.9	47.7	42.5
lstm_pe_1	PE	40.8	39.8	42.0	41.9
lstm_pe_2	PE	38.8	38.8	48.8	44.3

Table 3: Results for development and test corpora for the mono training case.

capture more tweets of the NEU category. This could be explained in part due to the different training strategy that focuses on giving the NEU tweets more weight.

It is also interesting to notice that for the three cross-lingual training cases, the best systems were SVMs. This could indicate that SVM is able to achieve good generalization even in the absence of variant-specific data.

6 Conclusions

We presented three approaches for TASS 2018 Task 1 about classifying the sentiment of tweets in different Spanish variants. The approaches we used are: SVM using word embedding centroids and manually crafted features, CNN using word embeddings as input, and LSTM using word embeddings, trained with focus on improving the recognition of neutral tweets. None of the classifiers was a clear winner in our experiments, as some of them worked better than others for different Spanish variants. However, we found that the training method used for the LSTMs significantly improved their macro-F measure by improving the detection of neutral tweets. In all cases, the use of word embeddings was key to improve the performance of the methods.

Submission	Var	Dev Acc	Dev F1	Test Acc	Test F1
svm_cross_es_1	ES	54.5	41.4	**57.2**	46.4
svm_cross_es_2	ES	53.2	41.4	55.5	**47.1**
cnn_cross_es_1	ES	48.8	41.9	52.4	45.0
cnn_cross_es_2	ES	52.8	41.0	56.3	44.8
lstm_cross_es_1	ES	47.2	41.0	49.8	43.8
lstm_cross_es_2	ES	43.7	41.6	46.6	47.0
svm_cross_cr_1	CR	58.0	43.2	**56.9**	**47.6**
svm_cross_cr_2	CR	58.7	45.8	54.2	47.4
cnn_cross_cr_1	CR	50.3	44.9	42.3	42.1
cnn_cross_cr_2	CR	55.0	44.8	55.1	46.2
lstm_cross_cr_1	CR	54.0	48.5	53.0	47.3
lstm_cross_cr_2	CR	45.3	42.0	46.8	44.4
svm_cross_pe_1	PE	39.0	31.4	50.5	44.4
svm_cross_pe_2	PE	40.2	32.0	**51.4**	**44.5**
cnn_cross_pe_1	PE	38.8	35.5	48.1	40.9
cnn_cross_pe_2	PE	40.6	38.6	43.8	39.1
lstm_cross_pe_1	PE	41.6	39.8	47.2	42.5
lstm_cross_pe_2	PE	42.2	41.9	46.5	44.4

Table 4: Results for development and test corpora for the cross training case.

		Predicted N	Predicted NEU	Predicted NONE	Predicted P
SVM	N	495	43	107	122
Real	NEU	81	19	51	65
	NONE	58	12	115	89
	P	83	14	65	480
CNN	N	608	24	38	97
Real	NEU	119	14	24	59
	NONE	112	9	73	80
	P	145	18	50	429
LSTM	N	437	138	118	74
Real	NEU	55	67	57	37
	NONE	28	62	133	51
	P	61	91	84	406

Table 5: Confusion matrix for the best system for each classifier family over the ES test corpus. The LSTM captures significantly more neutral tweets.

Bibliography

Azzinnari, A. y A. Martínez. 2016. Representación de Palabras en Espacios de Vectores. Proyecto de grado, Universidad de la República, Uruguay.

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En *RANLP*, páginas 50–54.

Cerón-Guzmán, J. A. 2016. Jacerong at tass 2016: An ensemble classifier for sentiment tweets at global level. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain.*

Cerón-Guzmán, J. A. 2017. Classifier ensembles that push the state-of-the-art in sentiment analysis of spanish tweets. En *Proceedings of TASS.*

Chollet, F. 2015. Keras. `https://github.com/fchollet/keras`.

Cruz, F. L., J. A. Troyano, B. Pontes, y F. J. Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Garcıa-Cumbreras, M., J. Villena-Román, E. Martınez-Cámara, M. Dıaz-Galiano, M. Martın-Valdivia, y L. na López. 2016. Overview of tass 2016. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain, September.*

Hurtado, L.-F., F. Pla, y J. González. 2017. ELiRF-UPV at TASS 2017: Sentiment Analysis in Twitter based on Deep Learning. En *Proceedings of TASS.*

Hurtado, L. F. y F. Plà. 2016. Elirf-upv en tass 2016: Análisis de sentimientos en twitter. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain.*

Kingma, D. P. y J. Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Martínez-Cámara, E., Y. Almeida Cruz, M. C. Díaz-Galiano, S. Estévez Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez Vázquez, A. Montejo Ráez, A. Montoyo Guijarro, R. Muñoz Guillena, A. Piad Morffis, y J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. En E. Martínez-Cámara Y. Almeida Cruz M. C. Díaz-Galiano S. Estévez Velarde M. A. García-Cumbreras M. García-Vega Y. Gutiérrez Vázquez A. Montejo Ráez A. Montoyo Guijarro R. Muñoz

Guillena A. Piad Morffis, y J. Villena-Román, editores, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volumen 2172 de *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Martínez-Cámara, E., M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, y J. Villena-Román. 2017. Overview of tass 2017. En J. Villena Román M. A. García Cumbreras D. G. M. C. Martínez-Cámara, Eugenio, y M. García Vega, editores, *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volumen 1896 de *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Moctezuma, D., M. Graff, S. Miranda-Jiménez, E. Tellez, A. Coronado, C. Sánchez, y J. Ortiz-Bejar. 2017. A genetic programming approach to sentiment analysis for twitter: Tass'17. En *Proceedings of TASS.*

Mohammad, S., F. Bravo-Marquez, M. Salameh, y S. Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. En *Proceedings of The 12th International Workshop on Semantic Evaluation*, páginas 1–17.

Montejo-Ráez, A. y M. C. Díaz-Galiano. 2016. Participación de SINAI en TASS 2016. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain.*

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

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, y E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Quirós, A., I. Segura-Bedmar, y P. Martínez. 2016. LABDA at the 2016 TASS challenge task: Using word embeddings for the sentiment analysis task. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain.*

Rosá, A., L. Chiruzzo, M. Etcheverry, y S. Castro. 2017. RETUYT en TASS 2017: Análisis de Sentimientos de Tweets en Español utilizando SVM y CNN. En *Proceedings of TASS.*

Saralegi, X. y I. San Vicente. 2013. Elhuyar at tass 2013. *XXIX Congreso de la Sociedad Espaola de Procesamiento de lenguaje natural, Workshop on Sentiment Analysis at SEPLN (TASS2013)*, páginas 143–150.

White, L., R. Togneri, W. Liu, y M. Bennamoun. 2015. How well sentence embeddings capture meaning. En *Proceedings of the 20th Australasian Document Computing Symposium*, página 9. ACM.

Wu, T.-F., C.-J. Lin, y R. C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005.