Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 109-116 recibido 02-12-2019 revisado 09-01-2020 aceptado 11-01-2020

Sentiment Analysis in Spanish Tweets: Some Experiments with Focus on Neutral Tweets

Análisis de Sentimiento para Tweets en Español: Algunos Experimentos con Foco en los Tweets Neutros

Luis Chiruzzo, Mathias Etcheverry, and Aiala Rosá
Universidad de la República, Montevideo, Uruguay
{luischir,mathiase,aialar}@fing.edu.uy

Abstract: We present different methods for Sentiment analysis in Spanish tweets: SVM based on word embeddings centroid for the tweet, CNN and LSTM. We analyze the results obtained using the corpora from the TASS sentiment analysis challenge, obtaining state of the art results in the performance of the classifiers. As the neutral category is the hardest one to classify, we focus in understanding the neutral tweets classification problems and we further analyze the composition of this class in order to extract insights on how to improve the classifiers.
Keywords: Sentiment Analysis, Spanish Tweets Analysis, Machine Learning, Deep Learning, Word Embeddings

Resumen: Presentamos diferentes métodos para análisis de sentimiento de tweets en español: SVM basado en centroide de word embeddings, CNN y LSTM. Analizamos los resultados otenidos usando el corpus de la competencia análisis de sentimiento TASS, obteniendo resultados en el estado del arte para nuestros clasificadores. Como la categoría de los neutros es la más difícil de clasificar, nos enfocamos en entender los problemas de clasificación de los neutros y analizamos la composición de esta clase en profundidad para obtener ideas sobre cómo mejorar los clasificadores.
Palabras clave: Análisis de Sentimiento, Análisis de Tweets en Español, Aprendizaje Automático, Aprendizaje Profundo, 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 shared 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].

These tasks have provided different resources available for research. The TASS workshop has generated several corpora for Sentiment Analysis in Spanish tweets, including different variants of Spanish. SemEval 2018 has also provided a corpus of Spanish tweets with polarities, but using a different set of classes than TASS.

Tweets present some special characteristics that must be taken into account: colloquial language, spelling errors, syntactic errors, abbreviations, use of symbols and URLs, lack of context, references to public events and personalities not explicitly mentioned, etc. Some of these problems can be easily overcome, but others, like syntactic errors or lack of context, have no simple solutions.

The most common approaches in recent years for tweets sentiment analysis are based on word embeddings and neural networks, to the detriment of tools for linguistic analysis, typically used to generate features for training ML algorithms, which had been applied in previous years. In many cases, information

[1] https://www.cs.york.ac.uk/semeval-2013/task2.html

[2] http://www.sepln.org/workshops/tass/2012/

ISSN 1135-5948. DOI 10.26342/2020-64-13

provided by subjective lexicons is still used.

During the TASS competitions, the measure used to rank the systems is the macro-F1 measure, which averages the F1 score for each class giving equal weight to all classes. Because of this, the participants tend to optimize their systems looking for the best macro-F1 score. However, the systems presented in the latest editions of the TASS workshop do not reach good results, thus leaving space for improvements. In particular, some papers report especially low results in neutral tweets classification. This could probably happen due to the unclear definition of these tweets, which frequently contain positive and negative fragments, resulting in a neutral global content; or due to the scant number of neutral tweets generally present in the training corpora. In any case, this low performance in the neutral class heavily penalizes the macro-F1 scores, even if the neutral tweets are not the most abundant class in the test corpus. One of the aims of this work is to try to analyze this neutral category in order to gain insights about its composition and the behavior of the classifiers around it.

In this paper we describe different approaches for Spanish tweet classification: an SVM-based classifier which uses a set of features, including word embeddings; and two deep neural network approaches: CNN and LSTM. We analyze the results obtained by each method, focusing on the improvement of neutral tweets classification.

The rest of this paper is organized as follows. Section 2 shows some background on this task and relevant related work. Section 3 describes the corpus we used and the preprocessing we made in order to work with it. Section 4 presents the linguistic resources we used to build our classifiers. Section 5 describes the three types of classifiers we used. Section 6 shows the results on the test corpus and analyzes the behavior of the classifiers, particularly over the neutral class. Finally, section 7 presents our conclusions and some future work.

2 Related Work

As in many NLP areas, in the last years most of the works on sentiment analysis have incorporated techniques based on Deep Learning and Word Embeddings, in search of improving results. In a review of Spanish Sentiment Analysis, covering works from 2012 to 2015 (Miranda and Guzmán, 2017), none of the described approaches is based on these methods, however, in recent editions of the TASS shared tasks (2017, 2018, 2019), the majority of participating systems rely on different neural network models and on the use of word embeddings (Manuel C. Díaz-Galiano, 2018; Martínez-Cámara et al., 2018; Díaz-Galiano et al., 2019; Martínez-Cámara et al., 2017). However, approaches based on classic machine learning models (like SVM), when including word embedding based features, remain competitive, reaching the top positions for some test corpora (Martínez-Cámara et al., 2018).

In TASS 2018 (task 1) three different corpora were available, each one for a different Spanish variant: Spain, Costa Rica, and Peru. The tagset used for tweets annotation included positive (`P`), negative (`N`), neutral (`NEU`) and no-opinion (`NONE`) classes. The best results were obtained by systems which used deep learning (Chiruzzo and Rosá, 2018; González, Pla, and Hurtado, 2018), SVM (Chiruzzo and Rosá, 2018), and genetic algorithms combined with SVM (Moctezuma, 2018). All of them used word embeddings for words and tweets representation. Results for monolingual experiments (using a single Spanish variant) were better than results for crosslingual experiments. As in previous TASS editions, neutral tweets are the most difficult to recognize.

SemEval 2018 (Mohammad et al., 2018) has included for the first time 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 (Convolutional Neural Networks and Recurrent Neural Networks) and SVM, based on word embeddings (Kuijper, Lenthe, and Noord, 2018; Rozental and Fleischer, 2018; Abdou, Kulmizev, and Ginés i Ametllé, 2018; González, Hurtado, and Pla, 2018). Some of them augmented the training set by translating English tweets (Kuijper, Lenthe, and Noord, 2018; Rozental and Fleischer, 2018). Other systems used subjective lexicons (Spanish lexicons and translated English lexicons).

3 Corpus

The experiments presented in this work are based on the corpora provided by different editions of the TASS sentiment analysis challenge (Martínez-Cámara et al., 2018; Martínez-Cámara et al., 2017).

We used three sets of corpora for Spanish variants spoken in different countries: Spain (ES), Costa Rica (CR) and Peru (PE), from the 2018 edition of TASS, and the general TASS training data from the 2017 edition of the competition. All the corpora are annotated with four possible polarity categories per tweet: `P`, `N`, `NEU` or `NONE`.

We joined the four sets to have a unique Spanish corpus, then divided in two subsets: training (90 %) and development (10 %). We used the test corpora distributed by TASS 2018 for evaluation. Table 1 shows the sizes of the different corpora and the number of tweets for each class.

Category	Train	Dev	Test
N	3227	361	1730
NEU	1109	123	747
NONE	2257	249	657
P	3607	400	1426
Total	10200	1133	4560

Tabla 1: Size and categories distribution for the different corpora

Each corpus was pre-processed as described in (Chiruzzo and Rosá, 2018; Rosá et al., 2017). We did not include any grammatical information, like lemma, POS-tag, morphological or syntactic information.

4 Resources

The following linguistic resources were used in our experiments:

- Subjective Lexicons: union of three subjective lexicons available for Spanish (Cruz et al., 2014; Saralegi and San Vicente, 2013; Brooke, Tofiloski, and Taboada, 2009).
- Word embeddings set: 300 dimension general purpose word embeddings set, trained by (Azzinnari and Martínez, 2016).
- Word Polarity Predictor: predictor trained using the subjective lexicons, taking a word vector as input and returning a real value for its polarity (Chiruzzo and Rosá, 2018).
- Category Markers: words that occur at least 75 % times in each category (Chiruzzo and Rosá, 2018).

5 Classifiers

This section describes the three approaches we used for classifying the polarity of Spanish tweets and gives details about the training of the methods.

5.1 SVM based approach

The SVM classifier configurations are almost the same as the ones described in (Rosá et al., 2017) and (Chiruzzo and Rosá, 2018), we used these features:

- Centroid of tweet word embeddings. (300 real values)
- Lexicon based Features:
  - Polarity of the nine (average length of tweets) more relevant words of the tweet according to the polarity predictor (those 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 and negative lexicons. (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)
- The thirty most relevant words from the training corpus, according to a bag of words (BoW) classifier. (30 boolean values)

This final attribute set was defined experimentally, evaluating on the development corpus. We started using just BoW attributes, obtaining better accuracy and macro-F1 as we increased the number of words (we selected the most relevant words for classifying the instances). This improvement stopped when we reached about a thousand words. We then included the centroid of the tweet word embeddings getting a better performance. The best results were obtained combining both types of attributes (BoW and the tweet centroid), increasing accuracy in 13 points and macro-F1 in 12 points. The optimal combination we found uses the centroid attributes plus the thirty most relevant words according to BoW.

The remaining attributes produced small improvements in the results as they were incorporated. While none of these attributes in particular provides a substantial improvement, the inclusion of all of them increased accuracy and macro-F1 in approximately 2 points, with respect to the results obtained using only the centroid and the thirty most relevant words. In particular, we found that the contribution of the subjective lexicon is not very relevant. Figure 1 shows different combinations of BoW and lexical features.

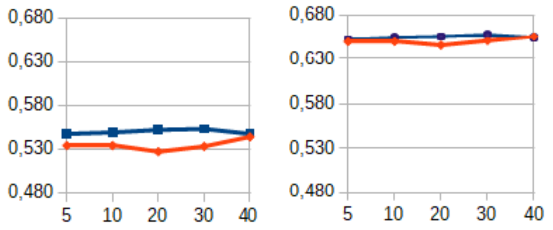

Figura 1: Macro-F1 (left) and Accuracy (right) for different combinations of BoW features, with lexical features (blue) and without lexical features (red)

In previous work, we included just five words from BoW features, in the current work, experiments on the development corpus showed that the best results are obtained using the thirty most relevant words from BoW features, most of which are words with a clear polarity, like *abrazo* (hug), *buen, buena, buenas, buenos* (different inflections of the word 'good'), *déficit* (deficit), *encanta* (love), *enhorabuena, felicidades* (variants of 'congratulations'), *feliz* (happy), *genial* (great), *gracias* (thanks), *impuestos* (taxes), *mejor* (better), *peor* (worse), *triste* (sad).

The SVM experiments were done using the *scikit-learn* toolkit (Pedregosa et al., 2011), applying the StandardScaler to the training dataset. We used the multiclass probability estimation method based on (Wu, Lin, and Weng, 2004) for training. In previous work, this method showed an improvement of 2.5 % in macro-F1 over the single class prediction.

5.2 CNN based approach

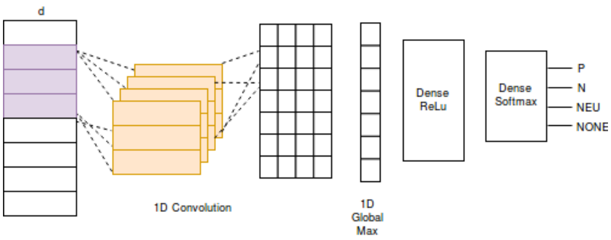

Figura 2: Overview of the convolutional neural network architecture

Convolutional Neural Networks are a class of neural networks invariant to the elements position in the input that are inspired by the cells found in the cats visual cortex. We consider two approaches using convolutional networks to perform tweets sentiment classification: 1) one branched CNN and 2) three branched CNN. The former consists in a standard CNN model, the input of the network is the sequence of word embeddings of dimension 300, corresponding to each word in the tweet, up to a maximum of 32 words. The input is fed to the 1D convolutional layer with 30 filters of dimension 256, then the output goes to a max pooling layer and a dense layer of dimension 200 with a dropout of 0,2, before going to the softmax layer for output. An overview of this architecture is shown in figure 2.

The three branched CNN is constituted by three convolutional sub-networks that consume the same input and process it independently. The three branched model considers two, three and four words of context, with 31, 30 and 29 filters of dimension 200, respectively. Then all outputs are concatenated and passed to a max pooling layer. Then, the max pooling output fed a dense layer of dimension 200 with a dropout of 0,2 before going to a softmax layer for output.

For training we keep a 70 %-30 % split for validation and use early stopping over the validation set for both networks.

5.3 LSTM based approach

Long Short-Term Memory neural networks take in consideration the whole sequence before yielding a result, they are a subclass of recurrent neural networks. In a way, what they do is calculate a sentence embedding, and then use that embedding to make a prediction. Our LSTM architecture uses the embedding for each word as input, up to a maximum of 32 words. This input is sent through a LSTM layer, for which we ignore the intermediate results, considering only the one after the whole sequence of words has been processed, we will call this output the sentence embedding. This sentence embedding of size 512 is then sent through a dense layer of size 200 with a dropout of 0.2, before getting the output through a softmax layer.

The initial experiments using this network yielded good accuracy results, but the macro-F1 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 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 these sets the *full* corpus and the *balanced* corpus.

The training strategy involves training one epoch with the full 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-F1 measure, as it captures a lot more tweets of the `NEU` category.

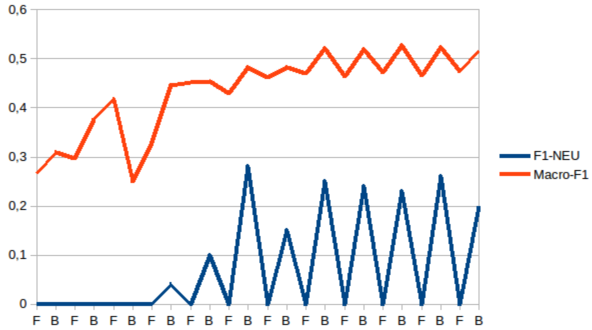

Figura 3: Macro-F1 and F1 for the `NEU` category during some iterations of the LSTM training. The iterations marked as F use the full corpus, iterations marked as B use the balanced corpus

When examining the training process more closely, we observed that after every full corpus iteration, the performance over the `NEU` category dropped to zero, while after every balanced corpus iteration the performance over that category improves. Stopping the training process after the balanced step yields better macro-F1 because it captures more `NEU` tweets. This is shown in figure 3.

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

6 Results

In this section we present the results for each of the classifiers trained only using training data or using training and development data, then we analyze the learning curve for the classifiers and finally we discuss the case of the tweets of the neutral category in more detail.

6.1 Results for classifiers

Table 2 shows the performance of the four classifiers trained using only training data and evaluating over the development and test sets. The SVM approach clearly outperforms the other classifiers in terms of accuracy, and on the development set is also the approach with highest macro-F1. Over the test set, however, both accuracy and macro-F1 drop significantly for all classifiers. In this case, the LSTM and the three-branched CNN approaches are the ones with the best macro-F1.

	Dev		Test	
Classifier	Acc	F1	Acc	F1
SVM	65.6	55.2	**56.2**	45.4
CNN1	55.3	49.7	48.9	44.7
CNN3	59.3	52.1	55.7	46.4
LSTM	55.9	52.9	49.9	**46.6**

Tabla 2: Results for development and test corpora training only with train data

When using the whole training and development sets for training and evaluating over the test set, as shown in table 3, the ranking between classifiers is similar but the figures change. The best accuracy is still achieved by the SVM approach and is almost the same as before, while the best macro-F1 is still achieved by the LSTM, but in this case the measure is improved by two points.

As can be seen in table 4, one of the reasons the LSTM could have gotten better

Classifier	Acc	F1
SVM	**56.3**	45.3
CNN1	52.2	43.7
CNN3	51.6	47.1
LSTM	52.1	**48.7**

Tabla 3: Results for test corpora training with train and development data

results over the test set was because it could 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. As we can see, the network captures more `NEU` tweets because it learned to predict a more balanced distribution. This strongly penalizes the accuracy, but is good for the macro-F1.

The three-branched CNN approach has the second best macro-F1 score, but when training with both training and development corpora its accuracy dropped respect to the other classifiers. The SVM approach is the most stable one, as its performance over the test set did not change significantly in the two models.

		Predicted			
		N	NEU	NONE	P
SVM	N	**1182**	55	258	235
Real	NEU	273	**49**	210	215
	NONE	157	14	**289**	197
	P	195	27	159	**1045**
CNN1	N	**1039**	151	244	296
Real	NEU	211	**97**	182	257
	NONE	153	50	**238**	216
	P	165	86	167	**1008**
CNN3	N	**1093**	319	222	96
Real	NEU	227	**242**	193	85
	NONE	142	131	**269**	115
	P	169	314	193	**750**
LSTM	N	**936**	489	184	121
Real	NEU	168	**305**	141	133
	NONE	97	180	**256**	124
	P	89	341	118	**878**

Tabla 4: Confusion matrix for the classifiers trained with training and development data. The LSTM captures significantly more neutral tweets

The accuracy and macro-F values we obtained are similar to the results reported in the TASS sentiment analysis challenge (Martínez-Cámara et al., 2018). We have slightly lower results than TASS monolingual experiments, and almost the same results as cross-lingual experiments. This is an expected result, since we used a cross-lingual corpus (composed of Spain, Costa Rica and Peru corpora) for all our experiments.

6.2 Learning curves

In order to see if using more annotated data would enable us to further improve our results, we experimented with varying corpus sizes, increasing the corpus size by 10 % and analyzing the behavior of the different classifiers at each step. In these experiments, we used the training and development sets together for training and we tested against the test corpus.

Figure 4 shows the macro-F1 measure for each classifier using different training sizes. The SVM is clearly improving linearly with the training set size, which indicates that using more data would keep improving this metric. The LSTM and CNN with one branch start to show better results using around 60 % of the corpus, and they also seem to keep improving given more data. The CNN with three branches seems to plateau around 80 % of the training corpus.

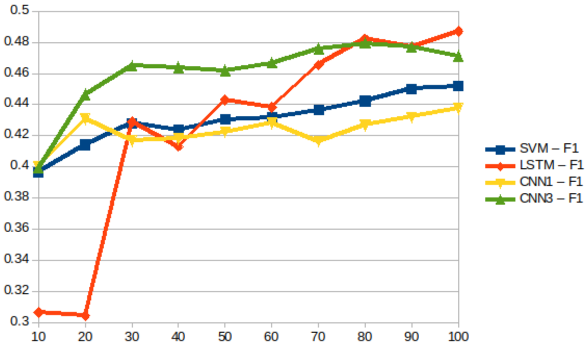

Figura 4: Macro-F1 of the classifiers trained with fractions of the corpus

6.3 Neutral tweets analysis

According to the guidelines defined for the annotation of the TASS corpus, the neutral class contains two very different subclasses: tweets without polarity, which we could call true neutrals (`NEU-NEU`), and tweets with positive parts and negative parts (`NEU-MIX`), in which none of the two polarities has a clear preponderance over the other. Two examples of these two subclases are:

- `NEU-NEU`: *@user se cosas de tiempos actuales* (@user I know things about current times)

- NEU-MIX: *feo es tener clases un feriado, por suerte yo no tengo* (it's ugly to have classes on a holiday, fortunately I don't)

While the overall results for neutral tweets are quite different for the different classifiers (as already stated, the LSTM-based system obtains a noticeably better result with this class than the other two), the three models behave in a similar way regarding the two subclasses of neutral tweets (CNN with one branch is not analyzed in this section).

Class	Real	SVM	CNN3	LSTM
NEU	747	7 %	30 %	41 %
NEU-NEU	276	4 %	29 %	31 %
NEU-MIX	471	8 %	31 %	48 %

Tabla 5: Percentage of correct predictions of Neutral tweets and subclasses NEU-NEU and NEU-MIX (training only with the train corpus)

	SVM	CNN3	LSTM
NEU-NEU vs NONE	121	101	85
NEU-NEU vs N	93	41	66
NEU-NEU vs P	50	54	40
NEU-MIX vs NONE	88	87	44
NEU-MIX vs N	191	113	124
NEU-MIX vs P	151	126	79

Tabla 6: NEU-NEU and NEU-MIX confusion with other classes

In all cases, the systems are better at recognizing tweets of the NEU-MIX category than those of the NEU-NEU category (see table 5). In addition, NEU-NEU tweets are confused by the three classifiers with the NONE class more frequently than with the N and P classes (see table 6). From a human point of view it is also difficult to distinguish the NEU-NEU class, which indicates that the tweet is subjective but with no polarity, from the NONE class, which indicates that the tweet transmits non-subjective information (for example, *Último recibo que pagaré será el de Setiembre. / The last receipt that I will pay will be the one of September*). On the other hand, the NEU-MIX tweets are often confused with the N and P classes. This behavior seems to be due to the fact that the NEU-MIX tweets contain portions with marked polarity (both positive and negative), while the NEU-NEU tweets should not contain a clear polarity at all.

7 Conclusion

We presented three approaches for classifying the sentiment of Spanish tweets. The approaches we used are: SVM using word embedding centroids and manually crafted features, one and three branched CNNs 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. However, we found that the training method used for the LSTM significantly improved its macro-F1 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.

Analyzing the learning curves of the classifiers, we concluded that most of them would still keep improving if there was more data available. However, it is clear that the bottleneck of the macro-F1 performance is still the neutral class.

We separated neutral tweets in two classes, NEU-MIX and NEU-NEU, and we analyzed classifiers mistakes on each class. We found that NEU-MIX tweets are usually confused with negative and positive classes, while NEU-NEU tweets are confused with the class NONE. We propose as future work to improve neutral tweets classification by segmenting the input into chunks with homogeneous sentiment polarity and feeding this chunks as units to the classifiers.

References

Abdou, M., A. Kulmizev, and J. Ginés i Ametllé. 2018. Affecthor at semeval-2018 task 1: A cross-linguistic approach to sentiment intensity quantification in tweets. In *12th International Workshop on Semantic Evaluation*, pages 210–217. ACL.

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

Brooke, J., M. Tofiloski, and M. Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, pages 50–54.

Chiruzzo, L. and A. Rosá. 2018. RETUYT-InCo at TASS 2018: Sentiment Analysis in Spanish Variants using Neural Networks and SVM. In *TASS 2018*.

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

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

Díaz-Galiano, M. C., M. García-Vega, E. Casasola, L. Chiruzzo, M. A. García-Cumbreras, E. Martínez-Cámara, D. Moctezuma, A. Montejo Ráez, M. A. Sobrevilla Cabezudo, E. Tellez, M. Graff, and S. Miranda. 2019. Overview of TASS 2019: One more further for the global spanish sentiment analysis corpus. In *IberLEF 2019*, Bilbao, Spain, September.

González, J., F. Pla, and L. Hurtado. 2018. ELiRF-UPV at TASS 2018: Sentiment Analysis in Twitter based on Deep Learning. In *TASS 2018*.

González, J.-Á., L.-F. Hurtado, and F. Pla. 2018. Elirf-upv at semeval-2018 tasks 1 and 3: Affect and irony detection in tweets. In *12th International Workshop on Semantic Evaluation*, pages 565–569. ACL.

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

Kuijper, M., M. Lenthe, and R. Noord. 2018. Ug18 at semeval-2018 task 1: Generating additional training data for predicting emotion intensity in spanish. In *12th International Workshop on Semantic Evaluation*, pages 279–285. ACL.

Manuel C. Díaz-Galiano, E. M.-C. y. M. Á. G. C. y. M. G. V. y. J. V. R. 2018. The democratization of deep learning in tass 2017. *Procesamiento del Lenguaje Natural*, 60(0):37–44.

Martínez-Cámara, E., Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez, A. Montejo Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, and V.-R. Julio. 2018. Overview of TASS 2018: Opinions, health and emotions. In *TASS 2018*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Martínez-Cámara, E., M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, and J. Villena-Román. 2017. Overview of tass 2017. In *TASS 2017*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Miranda, C. H. and J. Guzmán. 2017. A Review of Sentiment Analysis in Spanish. *Tecciencia*, 12:35 – 48, 06.

Moctezuma, D., O.-B. J. T.-E. M.-J. S. G. M. 2018. INGEOTEC solution for Task 1 in TASS'18 competition. In *TASS 2018*.

Mohammad, S., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *12th International Workshop on Semantic Evaluation*, pages 1–17.

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.

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

Rozental, A. and D. Fleischer. 2018. Amobee at semeval-2018 task 1: Gru neural network with a cnn attention mechanism for sentiment classification. In *12th International Workshop on Semantic Evaluation*, pages 218–225. ACL.

Saralegi, X. and 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)*, pages 143–150.

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