

# Stance and Gender Detection in Spanish Tweets

José-Ángel González, Lluís-F. Hurtado, Ferran Pla

Departament de Sistemes Informàtics i Computació  
Universitat Politècnica de València  
{jogonba2, lhurtado, fpla}@dsic.upv.es

**Abstract.** In this paper, we present a deep learning based system for the user profiling and stance detection tasks in Twitter. Stance detection consist of automatically determining from text whether the author is in favor of a given target, against this target, or whether neither inference is likely. The proposed system assembles Convolutional Neural Networks and Long Short-Term Memory neural networks. We use this system to address, with minor changes, both problems. We explore embeddings and one-hot vectors at character level to select the best tweet representation. We test our approach in the Stance and Gender Detection in Tweets on Catalan Independence track proposed at IberEval 2017 workshop. With the proposed approach, we achieve state-of-the-art results for the Stance detection subtask and the best results published until now for the Gender detection subtask.

**Keywords:** Stance detection, Gender detection, Deep learning, Twitter

## 1 Introduction

In recent years, Sentiment Analysis and Opinion Mining have aroused great interest in the natural language processing community. Sentiment Analysis consists of determining the polarity (positive, negative or none) expressed in a text. In addition, the use of social networks to express opinions on any topic has also become widespread. Both facts have facilitated that workshops on Sentiment Analysis on Twitter have attracted the interest of a large amount of research groups.

SemEval workshop has devoted several tasks to the analysis of sentiments on Twitter at different levels in which the English and Arabic languages have been involved [18], [19]. Also, for the Spanish language, within the conference of the SEPLN (Spanish Society for Natural Language Processing), TASS workshop has organized several tasks on Sentiment Analysis on Twitter in last years [5], [11].

However, to determine the stance of group of people about a certain target, it is not enough to know if the opinions are positive or negative. To determine which opinion is the majority, it is necessary to know if the users are in favor or against the target. Note that a positive opinion is different than an opinion in favor, given that you can express positive opinions but against a target; for

example, by praising the opposite position. Stance detection consist of automatically determining whether the author is in favor of the given target, against the target, or whether neither inference is likely. Sentiment Analysis can be done in a general way, but to carry out Stance detection properly, it is needed to know the target on which users express their opinions.

Moreover, knowing what type of users are participating in the discussion has great interest to conduct an adequate study of the opinions. In this regard, the goal of Author Profiling is to determine, from text, some characteristics of the author of that text. The most common characteristics that must be determined are gender and age.

Different international competitions have recently shown interest in Stance detection and Author Profiling: the task 6 at SemEval-2016 (Stance on Twitter)[14] that uses tweets in English and the Author Profiling task at PAN@CLEF 2016 [20] that uses texts written in English, Spanish, and Dutch extracted from several social networks.

Within the framework of the 33th conference of the SEPLN, it was organized the IberEval workshop that was dedicated to promoting the development of Human Language Technologies for Iberian languages. One of the share tasks proposed at IberEval was the Stance and Gender detection in Tweets on Catalan Independence (StanceCat)[23]. The aim of this task is to detect the author’s gender and his/her stance with respect to the ”independence of Catalonia” in tweets written in Spanish and Catalan.

In this paper, we present a system to address both Stance detection and User Profiling subtasks proposed at StanceCat task using the tweets written in Spanish. The system is based on the sequential representation of the tweets as input to a two-layer Convolutional Neural Network (CNN) [4] assembled with a final Long Short-Term Memory (LSTM) neural network [8]. Thus, it computes long-term relationships in the recurrent sub-network from short-term relationships computed in the convolutional sub-network.

A development phase was carried out to select the representation of the tweets that maximized the evaluation measure. The final representation was based on one-hot vectors at character level. Using the system described in this work, we achieve state-of-the-art results for the Stance detection subtask and the best results published until now for the Gender detection subtask for the Spanish language.

The rest of this paper is organized as follows. Section 2 presents StanceCat task and the corpus used in this paper. In Section 3, we present a short description of the system developed. Section 4 presents the experimental work conducted in this paper. Finally, in Section 5, we present some conclusions and the future work.

## 2 StanceCat at IberEval 2017

One of the tasks proposed in the 2017 edition of the IberEval workshop and in which more researchers participated was the StanceCat (Stance and Gender

detection in Tweets on Catalan Independence) task. This task had a double objective. On the one hand, to determine the gender of the author of the tweet and on the other hand to determine the stance of the author, expressed in the tweet, with respect to the independence of Catalonia.

The StanceCat organizers acquired a corpus of 10800 tweets -5400 written in Spanish and 5400 written in Catalan- published between September and December 2015. All the tweets include the hashtag *#Independencia* or *#27S* to ensure that the target was the Catalan Independence. The corpus is labeled in terms of the gender of the author of each tweet (MALE or FEMALE) and in terms of the stance of the author respect to the independence of Catalonia (FAVOR, AGAINST, or NONE).

Although corpus includes tweets in Spanish and Catalan, we have only used those written in Spanish language. Table 1 shows the number of samples per each label in the Spanish subset of the StanceCat corpus. Note that the corpus is unbalanced in terms of Stance detection, being a clear bias between AGAINST and NONE with respect to FAVOR, that only represents the 7.8% of the samples. However, this unbalance does not occur in the Gender detection subtask where both labels are equally represented both in the training and the test sets.

**Table 1.** Number of samples per label in the Spanish subset of the StanceCat corpus.

	Training set			Test set		
	MALE	FEMALE	Total	MALE	FEMALE	Total
FAVOR	190	145	335	48	36	84
AGAINST	753	693	1446	188	173	361
NONE	1216	1322	2538	305	331	636
Total	2159	2160	4319	541	540	1081

The official evaluation measure for the Stance detection subtask was the macro-average of  $F_1$ (FAVOR) and  $F_1$ (AGAINST), without taking into account the NONE label. For the Spanish subtask, a total of 31 runs were submitted by ten participating teams. The results ranged from 48.88 to 19.06. The best result was obtained by the *iTACOS*[10] team with 48.88 of macro-averaged  $F_1$ . They used Support Vector Machines (SVM) as classification paradigm. One of the aspect to be highlighted of the *iTACOS* proposal is the features selection process. They define three types of features: stylistic features, structural features, and context features including the text of the webs linked in the tweet.

Regarding the Gender detection subtask, the official evaluation measure was the Accuracy. For the Spanish subtask, a total of 19 runs were submitted by five participating teams. The results ranged from 68.55 to 47.64. The organizers proposed a baseline based on the Low Dimensionality Representation approach. The best result was achieved by the *ELiRF-UPV*[6] team with 68.55 of Accuracy which was the only one team that exceeded the baseline. The *ELiRF-UPV* team tested several classification paradigm but the best result was obtained when using SVM with bag-of-1grams and bag-of-2grams at character level.

More details about the task, the corpus, and the participants can be found in the review carried out by the task organizers and published in the IberEval workshop proceedings [23].

### 3 System Description

In this section we describe the main characteristics of the developed system to the Stance and Gender detection in Spanish tweets on Catalan Independence (StanceCat) task. This description includes the tweets preprocessing, their representation, and the system architecture.

#### 3.1 Preprocessing and Representation of the Tweets

As a previous step, we performed a preprocessing of the tweets. We removed the accents and converted all the text to lowercase. Web links and numbers were substituted by two generic tokens (URL and NUMBER, respectively). We maintained *hashtags*, *emoticons* and *user mentions* and they were not substituted by a generic label as we did in previous works.

We tested several sequential representations of the tweets to model the order among the different units considered. Specifically, we used the following approaches:

- Embeddings (at word level): we considered a tweet  $x$  as a sequence of words,  $x = x_1, x_2, \dots, x_n$  and we represented each word  $x_i$  by means of its embedding vector  $e(x_i) \in \mathbb{R}^{dw}$ , where  $dw$  is the dimension of the word embedding.
- One-hot vectors (at char level): we considered a tweet  $x$  as a sequence of characters,  $x = x_1, x_2, \dots, x_n$  and we represented each character  $x_i$  by means of a one hot vector,  $v(x_i) \in \mathbb{R}^{dc}$ , where  $dc$  is the number of different characters in the corpus. To generate the one hot vectors, we only considered the characters that appeared in the training set.

Using these two representations, we can contrast the semantic information provided by the embeddings against the stylistic characteristics modeled by means of the one-hot representation and select the more relevant representation to the Stance and Gender detection tasks.

Regarding embeddings, in previous works, we used Word2Vec models [12] [13] trained with the Spanish Wikipedia. In this work, we pre-trained our embeddings with 87 million tweets in Spanish. Our model is a skip-gram architecture. Each row of the lookup table has 300 dimensions and we used negative sampling as loss function.

Additionally, for the Stance detection task, we also used some emotion and polarity Spanish lexicons combined with the embeddings. In this case, each word  $x_i$  of a tweet was represented as a concatenation of its embedding  $e(x_i)$  with the information of the lexicons  $l(x_i)$ . Specifically, these lexicons were:

- ElHPolar: a list of 1889 positive and 3301 negative words created from different resources [21].

- ISOL: a list of 2509 positive and 5626 negative words developed from Bing Liu’s Opinion Lexicon [2]. This lexicon was automatically translated and manually reviewed [17].
- NRC Word-Emotion Association Lexicon: lexicon of 14182 words, automatically translated into Spanish. It contains information on polarity (negative, positive) and emotions (anger, fear, joy, sadness, disgust, trust, anticipation, and surprise) [16] [15].
- MLSenticon: a lexicon composed of 11542 words that provides a polarity estimate in the interval  $[-1, 1]$  [1].

### 3.2 System Architecture

We explored different models depending on the representation of the tweets. This way, CNN [4] assembled with LSTM [8] were used to deal with sequential representations of tweets. These models compute a representation based on long-term relationships in the recurrent sub-network from short-term relationships computed in the convolutional sub-network. Finally, using this representation, a fully connected single-layer network with softmax activation functions computes the outputs of the network.

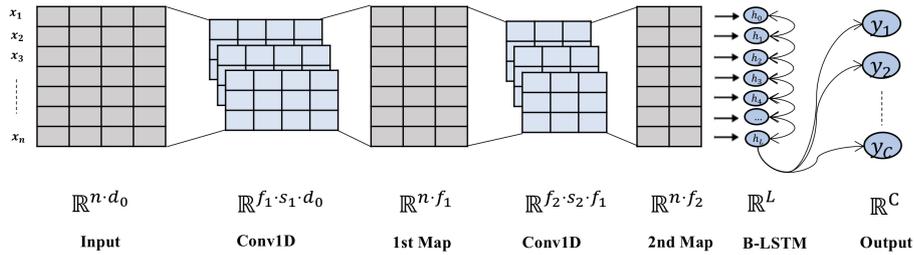


Fig. 1. System architecture.

Figure 1 shows a general scheme of the whole model, where  $n$  is the maximum number of elements in a tweet,  $d_0$  is the dimensionality of the representation of each element,  $f_i$  is the number of filters in the convolutional layer  $i$ ,  $s_i$  is the height of each filter in the layer  $i$ ,  $L$  is the dimensionality of the *output state* of the LSTM, and  $C$  is the number of classes of the task.

This model receives as input a tweet represented as a matrix  $M \in \mathbb{R}^{n \cdot d_0}$ , where  $d_0$  is denoted as  $dw$  at the word level, or  $dc$  at the character level. This input is directly passed to a CNN composed of two convolutional layers. For the Stance subtask, we set  $f_1 = 8$ ,  $f_2 = 16$  and  $s_1 = s_2 = 3$ . For the Gender detection subtask, these parameters were,  $f_1 = 128$ ,  $f_2 = 256$ ,  $s_1 = s_2 = 3$ . In both tasks padding was added in order to maintain the original number of rows.

A MaxPooling layer, with a  $size = 2$ , was applied to the output of the last convolutional layer. Later, a LSTM [8] was applied. For the Stance detection subtask we used an LSTM with  $L = 64$  and for Gender detection subtask we used a Bidirectional LSTM [22] with  $L = 2 * 128$ . Finally, in both cases, we used the last output of the LSTM as input to a fully connected layer of  $C$  neurons, whose function is to perform the classification task.

To speed up the convergence and favor a correct training of the models, we used BatchNormalization [9] after each layer (including the input to directly normalize the training data) for the Gender detection subtask. We used noisy representations of training data, obtained after applying a Gaussian noise layer with  $\sigma = 0.15$ , with the aim of improving generalization [7]. We used  $\tanh$  activation functions after each normalized output with BatchNormalization except in the last layer where a  $\text{softmax}$  activation function is used.

Finally, bucketing was used to train the models with variable length sequences. This allows us to build batches with representations of tweets that shares the same number of rows. In addition, the buckets were unsorted. This strategy was used both for reducing the training time and the over-fitting problem (similar to [3]).

## 4 Experimental Work

In this section we present the experimental evaluation of the systems introduced in Section 3. We report the results achieved both on the tuning and test phases.

### 4.1 Tuning Phase

In order to select the best representation and the best model for each task, a tuning process was performed. The training corpus provided by the organizers of the task was split into two sets, a set with the 80% of the tweets for training the model and the remaining 20% of the corpus was used as development set. These partitions were the same for all the tuning process.

Faced with the impossibility of testing all combinations of models and representations, only those combinations we thought that made more sense were considered.

**Table 2.** Tuning phase results for the Stance detection subtask.

Representation	$(F_{favor} + F_{against})/2$
Embeddings (Wikipedia)	51.84
Embeddings (Twitter)	52.19
Embeddings (Twitter) + Lexicons	51.08
One-hot (char level)	55.10

Table 2 shows the results on the development set for the Stance detection task. It can be observed how the embeddings, both of Wikipedia and Twitter,

behave worse than the One-hot representation. In turn, the addition of emotion/polarity information through lexicons worsens the results.

**Table 3.** Tuning phase results for the Gender detection subtask.

Representation	Accuracy (%)
Embeddings Twitter	70.13 $\pm$ 3.05
One-hot (char level)	80.90 $\pm$ 2.62

Table 3 shows the results on the development set for the Gender detection task. It can be seen that, also in this subtask, the *One-hot* representation is the one that obtains the best results.

## 4.2 Test Phase

Once the best representation of the tweets for each subtask was chosen, we tested the final models on the official test set.

Table 4 shows the results achieved by the proposed system compared with the best system at StanceCat competition.

**Table 4.** Results for the Stance detection subtask.

System	$(F_{favor} + F_{against})/2$
Proposed system	46.37
Best system at StanceCat[10]	48.88

Although we can not overcome the best results of the competition team, we achieved the third place in the competition. We carried out an study of the performance of our system at class level. Table 5 shows this analysis in terms of *Precision*, *Recall*, and  $F_1$  measure.

**Table 5.** Results at class level for Stance detection subtask.

Class	Precision	Recall	$F_1$
FAVOR	26.32	29.76	27.93
AGAINST	68.85	61.22	64.81
NONE	75.49	78.93	77.17

It can be seen that the worst results are obtained by the FAVOR class which is the class with minor number of samples, about 8% both in training and test sets. Our model did not include any mechanism to handle the imbalanced class problem. Note that the official evaluation measure do not take into account the NONE class in which we obtained the best results.

The results for Gender detection subtask, including confidence intervals, are shown in Table 6.

**Table 6.** Results for the Gender detection subtask.

System	Accuracy (%)
Proposed system	$81.03 \pm 2.33$
Best system at StanceCat[6]	$68.55 \pm 2.76$

We improved by 12.48 points the best previous result reported to this task [6] with the proposed architecture using one-hot vectors at character level as tweet representation. This improvement is statistically significant at 95% of confidence.

## 5 Conclusions and Future Work

In this paper, we presented a deep learning system for the Gender and Stance detection in Twitter. We explored different representation of the tweets -embeddings and one-hot vectors at character level- and an architecture that assembles CNN and LSTM neural networks.

We tested our approach in the Stance and Gender Detection in Tweets on Catalan Independence track at IberEval 2017 workshop. With the proposed approach, we achieved state-of-the-art results for the Stance detection subtask and the best results published until now for the Gender detection subtask.

As future work, we plan to carry out a study of the obtained results. With this study, we will try to answer questions such as, why representations at the character level are more relevant than those based on embeddings for the addressed tasks? We want to work in the development of techniques to handle the imbalanced classes, which has proved to be a very important problem in the Stance detection task.

We also plan to work with the subset of the corpus written in Catalan to verify the usefulness of the proposed system.

## Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER funds under projects ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics (TIN2014-54288-C4-3-R); and AMIC: Affective Multimedia Analytics with Inclusive and Natural Communication (TIN2017-85854-C4-2-R).

The authors thank the organizers of the *Stance and Gender Detection in Tweets on Catalan Independence* track for provide us the StanceCat corpus.

## References

1. Cruz, F.L., Troyano, J.A., Pontes, B., Ortega, F.J.: Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications* 41(13), 5984 – 5994 (2014), <http://www.sciencedirect.com/science/article/pii/S0957417414001997>
2. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. pp. 231–240. WSDM '08, ACM, New York, NY, USA (2008), <http://doi.acm.org/10.1145/1341531.1341561>
3. Doetsch, P., Golik, P., Ney, H.: A comprehensive study of batch construction strategies for recurrent neural networks in mxnet. *CoRR abs/1705.02414* (2017), <http://arxiv.org/abs/1705.02414>
4. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36(4), 193–202 (Apr 1980), <https://doi.org/10.1007/BF00344251>
5. García Cumbreiras, M.A., Villena Román, J., Martínez Cámara, E., Díaz Galiano, M.C., Martín Valdivia, M.T., Ureña López, L.A.: Overview of TASS 2016. In: *Proceedings of TASS 2016*. pp. 13–21. CEUR Workshop Proceedings. CEUR-WS.org (2016)
6. González, J.A., Pla, F., Hurtado, L.F.: ELiRF-UPV at IberEval 2017: Stance and Gender Detection in Tweets. In: *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*. pp. 193–198. IberEval 2017, CEUR Workshop Proceedings. CEUR-WS.org (2017)
7. Grandvalet, Y., Canu, S., Boucheron, S.: Noise Injection: Theoretical Prospects. *Neural Computation* 9(5), 1093–1108 (1997), <http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.5.1093>
8. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. *Neural Computation* 9(8), 1735–1780 (1997)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR abs/1502.03167* (2015), <http://arxiv.org/abs/1502.03167>
10. Lai, M., Cignarella, A.T., Hernández Faras, D.I.: iTACOS at IberEval2017: Detecting Stance in Catalan and Spanish Tweets. In: *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*. pp. 185–192. IberEval 2017, CEUR Workshop Proceedings. CEUR-WS.org (2017)
11. Martínez Cámara, E., Díaz Galiano, M.C., García Cumbreiras, M.A., Vega, M.G.: Overview of tass 2017. In: *Proceedings of TASS 2017*. pp. 13–21. CEUR Workshop Proceedings. CEUR-WS.org (2017)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *CoRR abs/1301.3781* (2013), <http://arxiv.org/abs/1301.3781>
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. *CoRR abs/1310.4546* (2013), <http://arxiv.org/abs/1310.4546>
14. Mohammad, S.M., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: *Proceedings of the International Workshop on Semantic Evaluation. SemEval '16, San Diego, California (June 2016)*

15. Mohammad, S.M., Turney, P.D.: Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. pp. 26–34. CAAGET '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), <http://dl.acm.org/citation.cfm?id=1860631.1860635>
16. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon 29(3), 436–465 (2013)
17. Molina-González, M.D., Martínez-Cmara, E., Martín-Valdivia, M.T., Perea-Ortega, J.M.: Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications* 40(18), 7250 – 7257 (2013), <http://www.sciencedirect.com/science/article/pii/S0957417413004752>
18. Nakov, P., Ritter, A., Rosenthal, S., Stoyanov, V., Sebastiani, F.: SemEval-2016 task 4: Sentiment analysis in Twitter. In: Proceedings of the 10th International Workshop on Semantic Evaluation. SemEval '16, Association for Computational Linguistics, San Diego, California (June 2016)
19. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 Task 4: Sentiment Analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation. SemEval '17, Association for Computational Linguistics, Vancouver, Canada (August 2017)
20. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16, pp. 332–350. Springer International Publishing (2016)
21. Saralegi, X., San Vicente, I.: Elhuyar at tass 2013. In: XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural, Workshop on Sentiment Analysis at SEPLN (TASS2013). pp. 143–150 (2013)
22. Schuster, M., Paliwal, K.: Bidirectional recurrent neural networks. *Trans. Sig. Proc.* 45(11), 2673–2681 (Nov 1997), <http://dx.doi.org/10.1109/78.650093>
23. Taulé, M., Martí, M., Rangel, F., Rosso, P., Bosco, C., Patti, V.: Overview of the task of Stance and Gender Detection in Tweets on Catalan Independence at IBEREVAL 2017. In: Martínez, R., Gonzalo, J., Rosso, P., Montalvo, S., de Albornoz, J.C. (eds.) Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL). pp. 157–177. CEUR Workshop Proceedings. CEUR-WS.org, 2017, Murcia (Spain) (September 2017)