ELiRF-UPV at IberEval 2017: Stance and Gender Detection in Tweets

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

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

Abstract. This paper describes the participation of ELiRF-UPV team at the two Spanish subtasks of the Stance and Gender Detection in Tweets on Catalan Independence track of the IberEval workshop. We tested several approaches based on different models and tweet representations. Our best approaches are based on neural networks with one-hot vector representation and Support Vector Machines using bag-of-ngrams of chars.
We achieved the first place on the gender detection subtask and the fourth place on the stance detection subtask.

Keywords: Neural Networks, Support Vector Machine, bag-of-words, one-hot vectors

1 Introduction

Stance detection consist of automatically determining from text whether the author is in favor of the given target, against the given target, or whether neither inference is likely.

Different international competitions have recently shown interest in these subjects: Stance on Twitter, task 6 at SemEval-2016 [5] and Gender detection at PAN@CLEF 2016 [8].

Stance and Gender detection in Tweets on Catalan Independence is one of the tracks proposed at Ibereval 2017 workshop [9]. The aim of this task is to detect the author's gender and stance with respect to the target "independence of Catalonia" in tweets written in Spanish and/or Catalan.

2 Corpus Description

The corpus is composed by tweets labeled with respect to the independence of Catalonia (three classes: AGAINST, NEUTRAL, FAVOR) and with respect to the gender of the author of each tweet (two classes: MALE and FEMALE).

These tweets are provided in Spanish and Catalan, however, we have only worked with the Spanish version of the proposed corpus. On the other hand, it is necessary to take into account that the corpus is unbalanced in terms of stance

detection, being a clear bias between classes AGAINST and NEUTRAL with respect to class FAVOR. This unbalance does not occur in the gender detection subtask as can be seen in Table 1.

Table 1. Number of samples per class in the Spanish subset of the corpus.

	Male	Female	Total
Against	753	693	1446
Neutral	1216	1322	2538
Favor	190	145	335
Total	2159	2160	4319

3 System Description

In this section we describe the main characteristics of the system developed to the Stance and Gender Detection in Tweets on Catalan Independence track of the IberEval workshop. This description includes the preprocessing used, the different tweets representations used and, the different models that were taken into account during the tuning phase.

3.1 Preprocessing

The preprocessing process of the tweets was a bit different depending on the subtasks. In both cases, we removed the accents and converted all the text to lowercase. The web links (URL), and the numbers were substituted by a specific label.

We assumed that the hashtags, the emoticons and the mentions to other users would be informative to determine the opinions of a user but not his/her gender. Accordingly to this assumption, we substituted the hashtags, the emoticons and user's mentions by a specific label for the Gender subtask, but we kept their values for the Stance subtask.

3.2 Tweets representation

We considered different approaches to represent the tweets:

- Embeddings. Sequential representations of words represented with embeddings Word2Vec [3], [4] [7] learned from the Spanish version of Wikipedia [10].
- Bag-of-ngrams. We tested as features, unigrams and bigrams of words and chars using a bag-of-ngrams representation.
- One-hot vectors. We also tested unigrams and bigrams of chars using a one-hot vector representation.

Since stance detection may be related, in some way, to sentiment analysis, we tested the use of polarity lexicons for the Stance subtask. Specifically, we tried to include NRC lexicon [6] as extra features for stance detection.

3.3   Models

We explored different models depending on the representation of the tweets. This way, Long short-term memory (LSTM) neural networks [2] assembled with Convolutional neural networks (CNN) were used to deal with the sequential representations (embeddings) and the one-hot vectors of chars representation, while for the bag-of-ngrams (both at the word level and at the char level) representation Support Vector Machines (SVM) with linear kernel and Multilayer Perceptrons (MLP) were used.

The NRC polarity lexicon was used only with the embeddings representation. The topology used was similar to the one described in [1] but without the subnet dedicated to processing the sequences formed with embeddings obtained from the training corpus.

3.4   Tuning

In order to select the representations and the models (including their parameters) more appropriated to each subtask, a tuning process was performed. The corpus provided by the organizers of the task was split into two sets, a set with the 80% of the tweet for learning the model and the remaining 20% of the corpus was used as tuning. The partitions were the same for all the tuning process. For tuning the models of each subtask, the official evaluation measure of each subtask was taken into account as optimization criterion.

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 shows the most relevant combinations of features and models as well as the results obtained during the tuning phase.

Table 2. Results obtained in the tuning phase.

System	Features	Stance (F_1)	Gender (Acc.)
CNN+LSTM	Embeddings	51.84	64.47%
CNN+LSTM	Emb+NRC	48.80	-
CNN+LSTM	One-Hot	**55.10**	-
MLP	Word 1-2grams	-	59.72%
MLP	Char 2-grams	-	63.81%
SVM	Word 1-2grams	-	58.30%
SVM	Char 2-grams	-	**66.92%**
SVM	Char 1-2-grams	-	**66.99%**

Regarding the stance detection subtask, as can be seen in the first row of Table 2, with sequential representation at word level (Wikipedia embeddings)

the result obtained was 51.84 for F_1 measure. Unfortunately, worse results were obtained when the polarity sequence according to the NRC lexicon was added to the embeddings representation (Emb+NRC).

We also tried the use of a different sequential representation formed by one-hot vectors at character level and processed with the same network used in the previous experiments. This representation together with CNN + LSTM obtained the best results on tuning, 55.10 of F_1 measure as can be seen in the third row of Table 2.

Regarding the gender detection subtask, although the sequential representation at word level (embeddings from Wikipedia) processed using CNN + LSTM obtained good results, 64.47% in Accuracy, the best results in the tuning phase were achieved by the representations based on bag-of-ngrams of chars. The representation based on bag-of-ngrams of words achieved significantly worst results.

The models that obtained better results were the Support Vector Machines with linear kernel. Specifically, the SVM model using bag-of-unigram of chars as representation of the tweets achieved 66.92% of Accuracy; while adding bag-of-bigrams of chars to the previous model slightly increases the Accuracy to 66.99%. These results correspond to the last two rows in Table 2.

4 Results

In view of the results obtained during the tuning phase and due the limitation of the track, we decided to send the following two runs to the competition.

- run1
    - stance detection: CNN + LSTM + char-one-hot
    - gender detection: SVM + bag-of-2grams of chars
- run2
    - stance detection: CNN + LSTM + char-one-hot (the same as in run1)
    - gender detection: SVM + bag-of-1grams of chars + bag-of-2grams of chars (the best accuracy at tunning)

Table 3 and Table 4 show the official result obtained by our systems in the stance detection subtask and the gender detection subtask respectively. The position obtained by our system in the competition is also included in parenthesis.

Table 3. Oficial results for the Stance detection subtask.

run	System and Features	F_1
run1/run2	CNN+LSTM + char One-hot vectors	46.37 (4)

Once we have analyzed the results, both in the tuning phase and in the official competition, we want to point out some interesting things.

In both subtasks, methods based on deep-learning have shown to offer competitive results. However, in the case of gender detection subtask, the best results

Table 4. Oficial results for the Gender detection subtask.

run	System and Features	Acc.
run1	SVM + bag-of-2grams of chars	68.55% (1)
run2	SVM + bag-of-1grams + bag-of-2grams of chars	58.74% (14)

have been obtained with a priori simple model, SVM and bag-of-chars. We hypothesize that the good results achieved by SVM models in this subtask is due to the greater robustness of these models (compared with deep-learning based models) to deal with the bias problem. If the imbalance is very large, it can cause that the network assigns all the samples only to the majority classes. The solution used for the stance detection subclass was to perform a scaling of the loss function during the training phase. This has prevented the network from classifying all tweets in the AGAINST and NEUTRAL classes (the majority classes with much difference in the Spanish version of the stance detection corpus).

Regarding the stance detection subtask, a sequential character-level representation has been chosen due to the increasing interest this kind of representations are having in the deep-learning area and the good results they are achieving [11]. In this way, we have been able to verify that, effectively, this type of representations (in conjunction with neural networks that handle sequences) provide competitive results in text classification tasks such as the stance detection subtask.

5   Conclusions and Future work

We have presented the participation of the ELiRF-UPV team at the Stance and Gender Detection in Tweets on Catalan Independence track of the IberEval workshop. Our team has participated in the two Spanish subtasks of the track and has achieved competitive results. Our best approaches were based on neural networks with sequential representation of the tweets and Support Vector Machines with bag-of-ngrams of chars.

As future work, we plan to use representations based on one-hot vectors at character level and CNN + LSTM on other tweet classification problems (TASS, SemEval, ...) in order to study their behavior in tasks other than stance detection.

Acknowledgements

References

1. González, J.A., Pla, F., Hurtado, L.F.: ELiRF-UPV at SemEval-2017 Task 4: Sentiment Analysis using Deep Learning. In: Proceedings of the 11th International Workshop on Semantic Evaluation. pp. 722–726. SemEval '17, Association for Computational Linguistics, Vancouver, Canada (August 2017)
2. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735–1780 (1997)
3. 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`
4. 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`
5. 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)
6. Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon 29(3), 436–465 (2013)
7. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), `http://is.muni.cz/publication/884893/en`
8. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16, pp. 332–350. Springer International Publishing (2016)
9. 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: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017). CEUR Workshop Proceedings. CEUR-WS.org, 2017, Murcia (Spain) (September 2017)
10. Wikipedia: Wikipedia spanish dumps (2017), `https://dumps.wikimedia.org/eswiki/`, [Online; accessed 18-May-2017]
11. Zhang, X., Zhao, J., LeCun, Y.: Character-level Convolutional Networks for Text Classification. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. pp. 649–657. NIPS'15, MIT Press, Cambridge, MA, USA (2015)