Shared Task on Stance and Gender Detection in Tweets on Catalan Independence - LaSTUS System Description

Francesco Barbieri
francesco.barbieri@upf.edu

Universitat Pompeu Fabra, Barcelona, Spain

Resumen In this paper we describe the system LaSTUS presented in the shared task on Stance and Gender Detection in Tweets on Catalan Independence, in the context of IberEval 2017. We participated to the task using FastText a linear model, extension of the classic bag of word. We also use pre-trained embeddings trained on 5 million tweets posted in Spain.

1. Introduction

In the past few years the debate on Catalan independence has been quite discussed in politics. The topic generated a lot of discussion as well in social media. In the shared task "Stance and Gender Detection in Tweets on Catalan Independence"[8] the organizers proposed a task to automatically recognize if a document (a tweet) is in favor or against the Catalan independence. Such automatic systems are very useful in practice, in order to analyze people opinion about a specific topic [7]. To successfully detect stance, automatic systems need to identify important bits of information that may not be present in the focus text. Moreover, this task is harder then the classic Sentiment Analysis task, since understanding whether the polarity of the tweet is positive or negative is not sufficient to understand the opinion of the author of the tweet.

The shared task also included a gender identification challenge, in order to study the demographic of the debate. The documents were in Spanish and Catalan. In the next section we will describe the tasks and the dataset provided by the organizers. In Section 3 we describe the system we used, and in Section 4 we show the results of our system.

2. Task and Dataset

The shared task included two tasks (for Spanish and Catalan tweets) [8]:

1. **Stance Detection**: *Given a message, decide the stance taken towards the target Çatalan Independence". The possible stance labels are: FAVOR, AGAINST and NONE.*
2. **Identification of Gender**: *Given a message, determine its author's gender. The possible gender labels are: FEMALE and MALE.*

The dataset [3] used in the tasks included tweets retrieved during the regional elections in September 2015, and the political debate was focused on a possible independence of Catalonia. The dataset included 8638 tweets for the stance and for the gender recognition tasks (4319 in Spanish and 4319 in Catalan). In Table1 we report examples from the dataset.

Spanish		
T1	**F**	Lo dije ayer y lo repito: votar algo que no sea 'Junts pel Si' o la CUP es tirar el voto a la basura. #somriureCUP
	N	Primeros datos de participación. 34,78 %. Un 5 % más a estas horas que en 2012 #27S
	A	#27S ¡Sí! ¡Soy ESPAÑOL!
T2	**M**	Artur Mas llamando a todos sus colegas empresarios, le falta un 3 % para llegar al 50 %. #27S
	F	En unas plebiscitarias (votas una preguntas binaria) ¿prevalecen votos (ciudadanos) o escaños? #27S
Catalan		
T1	**F**	Avui #si ha arribat el dia #27S serà un gran dia. Gràcies a tothom que hi ha treballat tant per fer-ho possible
	N	A #Sants n'hi ha que van a votar preparats #27S
	A	A casa hem jugat a les votacions i ma filla diu q ha votat al #presidentMas :(#epicfail #27S
T2	**M**	Avui farem història ?????????? #27s
	F	Bon dia Catalunya! Llibertat i democràcia. Cap a omplir les urnes! #27S

Cuadro 1: Examples of the dataset for each language and label of the two tasks. T1 is the stance detection task (Favor, Neutral, Against) and and T2 is the gender identification Task (Male and Female).

In addition to these tweets we also use a corpus o 5 million tweets posted in Spain between October 2015 and December 2016 in Spain, in order to train pre-trained vectors.

3. Our System

In this section we will describe the system we presented to the shared task. In the first sub-section we describe the preprossing pipeline, and in the second sub-section we describe the FastText classifier.

3.1. Preprocessing

Tweet texts were preprocessed with a modified version of the CMU Tweet Twokenizer [4], where we changed several regular expressions and added a Twitter emojis

vocabulary to better tokenize the tweets[1]. We also removed, from each tweet, all hyperlinks, and lowercased all textual content in order to reduce noise and sparsity. We also replace each user mention with the token "@user".

3.2. FastText

Fastext[2] [5] is a linear model for text classification. We decided to employ FastText as it has been shown that on specific classification tasks, it can achieve competitive results, comparable to complex neural classifiers (RNNs and CNNs). The best feature of FastText is the speed as it can be much faster than complex neural models. The FastText algorithm is similar to the CBOW algorithm [6], where the middle word is replaced by the label. Given a set of N documents, the loss that the model attempts to minimize is the negative log-likelihood over the labels:

$$loss = -\frac{1}{N}\sum_{N}^{n=1} e_n \log(softmax(BA_{x_n}))$$

where e_n is the label included in the n-th tweet, represented as hot vector. A and B are affine transformations (weight matrices), and x_n is the unit vector of the bag of features of the n-th document (comment). The bag of features is the average of the input words, represented as vectors with a look-up table.

We initialize the look-up table with pre-trained embeddings trained with the algorithm of [2], an extension of the continuous skipgram algotithm [6], where also the subinformation of the words is taken in account (by representing each word with a bag of n-grams, i.e. the sum of the vector representation of each n-gram included in the word). We pre-train the vectors on 5 million tweets geo-localized in Spain (see Section 2).

4. Results and Discussion

In this section we show the results of the model in the shared task and discuss them. In Table 2 are reported the results for the two tasks in the two languages. We show results of the best participant model, our model described in the previous section and also the ranking position of our model (comparing to other participant models).

In Table 2 we can see that our model is somehow competitive in the Stance-ES task and Gender-CA where it is outperformed by the best systems of four points. In the other two tasks (Stance-CA and Gender-ES) our model performs quite poorly comparing to the best system (8 points difference). We are not aware of the models used by other participants and can not infer the reason of these results. We can not even say that our system is better in one language or in one task as our best results are in Stance-ES and Gender-CA.

We believe that one of the problem of our system was the preprocessing: removing the user mentions (user) was not a good idea, as the user mentions could include important insights about the stance of the tweet. Also, we need to explore whether our

[1] http://www.ark.cs.cmu.edu/TweetNLP/

[2] https://github.com/facebookresearch/fastText

	Stance		Gender	
	ES	**CA**	**ES**	**CA**
Best Model	0.49	0.49	0.69	0.49
Our Model	0.46	0.40	0.61	0.44
Ranking	4 of 10	6 of 9	4 of 5	3 of 4

Cuadro 2: Results of the best model and our model in the two tasks and two languages. The Macro F1 is used in the Stance results and for the Gender task the accuracy. It is also reported the ranking of our system compared to other participant.

system was overfitting the training dataset, as using systems like Bag of Words, or similar methods, can lead to model a specific topic instead of modeling the target labels [1].

5. Conclusions

In this paper we describe the system we presented at the shared task on Stance and Gender Detection in Tweets on Catalan Independence. We used the FastText classifier with pre-trained embeddings trained on 5 million tweets. Our model performances are acceptable in some tasks, but in other tasks are very poor, suggesting that we need to improve the system. We look forward to see how other participants tackled the problem of Stance and Gender classification.

Referencias

1. Barbieri, F., Ronzano, F., Saggion, H.: How topic biases your results? a case study of sentiment analysis and irony detection in italian. In: Recent Advances in Natural Language Processing, RANLP. pp. 41–47. Bulgaria (2015)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
3. C. Bosco, M. Lai, V.P.F.R.P.R.: Tweeting in the debate about catalan elections. Language Resources and Evaluation Conference (LREC), workshop on Emotion and Sentiment Analysis Workshop (ESA) (2016)
4. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 42–47. Association for Computational Linguistics (2011)
5. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 2017 Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Valencia, Spain (April 2017)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

7. 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)
8. Taulé, M., Martí, M.A., 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: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL). CEUR Workshop Proceedings, CEUR-WS.org, Murcia, Spain (September 2017)