Neural models for StanceCat shared task at IberEval 2017

Luca Ambrosini[1] and Giancarlo Nicolò[2]

[1] Scuola Universitaria Professionale della Svizzera Italiana
[2] Univesitat Politècnica De València
luca.ambrosini@supsi.ch
giani1@inf.upv.es

Abstract. This paper describes our participation in the *Stance and Gender Detection in Tweets on Catalan Independence (StanceCat)* task at IberEval 2017. Our approach was focused on neural models, firstly using classical and specific model from state of the art, then we introduce a new topology of convolutional network for text classification.

1 Introduction

The raising of social networks as worldwide means of communication and expression, is gaining lot of interest from company and academia, due to the huge availability of daily contents published by users. Focusing on academia perspective, especially in the Natural Language Processing field, the contents available in form of written text are really useful for the study of specific open problems, where the stance detection related to political events is an example, and the *Stance and Gender Detection in Tweets on Catalan Independence (StanceCat)* task at IberEval 2017 is a concrete application.
In *StanceCat*, the principal aim is to automatically detect if the text's author is in favor of, against, or neutral towards the Catalan Independence. Moreover, as a secondary aim, participants are asked to infer the author's gender.
To tackle the described problem we built a *stance&gender detection* system mainly decomposed in two modules: text pre-processing and classification model. During the system's tuning process, different design choices were explored trying to find the best modules' combination and from their anlysis some interesting insight can be drawn.
In the following sections we firstly describe the StanceCat task (Section 2), then we illustrate the module's design of developed stance&gender detection system (Section 3), after that, an evaluation of the tuning process for submitted systems is analysed (Section 4), finally, conclusion over the whole work are outlined (Section 5).

2 Task definition

The StanceCat shared task aim was to detect the author's gender and stance with respect to the target *independence of Catalonia* in tweets written in Spanish and/or Catalan, where participants is allowed in the detection of both stance and gender or only in stance detection.
Participants had access to a labelled corpus of 4319 tweets for each language. We analysed it and find the following statistical informations presented in tables 1 and 2.

3 Systems description

In this section we describe the stance&gender detection systems. Organizing the system by modules, it can be divided in two blocks: text pre-preprocessing (Section 3.1) and classification model (Section 3.2).

Label	Favor	Neutral	Against	Total
ES	335	2538	1446	4319
CA	2648	1540	131	4319

Table 1: Statistical analysis of tweets' label from given corpus.

Tweets	Average	Deviation	Max
ES	14	3	23
CA	13	4	20

Table 2: Statistical analysis of given corpus' tweets regarding words length.

3.1 Text pre-processing

Regarding the text pre-preprocessing, has to be mentioned that the corpus under observation can not be treated as proper written language, because computer-mediated communication (CMC) is highly informal, affecting diamesic[3] variation with creation of new items supposed to pertain lexicon and graphematic domains [6,7]. Therefore, our pre-processing follows two approaches: classic and microblogging related. As classic aproach we used stemming (i.e., ST), stopwords (i.e., SW) and punctuation removal (i.e., PR). For microblogging approach we focus our attention over the following items: (i) mentions (i.e., MT), (ii) smiley (i.e., SM), (iii) emoji (i.e., EM), (iv) hashtags (i.e., HT), (v) numbers (i.e., NUM), (vi) URL (i.e., URL) (vii) and Tweeter reserve-word as RT and FAV (i.e., RW). For each of these items we leave the possibility to be removed or substituted by constant string.
In relation to above approaches we implement them using the following tools: (i) NLTK [3] and (ii) Preprocessor[4].

3.2 Classification models

Following, we describe the neural models used for the classification module. Before introducing the models we describe the specific text representation used as input layer Section 3.2 (i.e., sentence-matrix).

Text representation To represent the text we used word embeddings as described by [4], where *words* are represented as vectors of real number with fixed dimension $|v|$. In this way a whole sentence s, with length $|s|$ its number of word, is represented as a *sentence-matrix* M of dimension $|M| = |s| \times |v|$. $|M|$ has to be fixed a priori, therefore $|s|$ and $|v|$ have to be estimated. $|v|$ was fixed to 300 following [4]. $|s|$ was estimated analyzing table 2, in details we decided to fix it as the sum of average length plus the standard deviation (i.e. $|s| = 17$ for both language), with this choice input sentences longer than $|s|$ are truncated, while shorter ones are padded with null vectors (i.e., a vector of all zeros).
Choosing words as elements to be mapped by the embedding function, raise some challenge over the function estimation related to data availability. In our case the available corpus is very small and estimated embeddings could lead to low performance. To solve this problem, we decided to used a pre-trained embeddings estimated over Wikipedia using a particular approach called *fastText* [4].

[3] The variation in a language across medium of communication (e.g. Spanish over the phone versus Spanish over email)
[4] Preprocessor is a preprocessing library for tweet data written in Python, https://github.com/s/preprocessor

Convolutional Neural Network. Convolutional Neural Networks (CNN) are considered state of the art in many text classification problem. Therefore, we decide to use them in a simple architecture composed by a convolutional layer, followed by a *Global Max Pooling* layer and two dense layers.

Dilated KIM. This model is our new topology of CNN. It can be seen as an extension of Kim's model [1] using the dilation ideas from computer graphics field [13].
The original Kim's model is a particular CNN where the convolutional layer has multiple filter widths and feature maps. The complete architecture is illustrated in Figure 1, here the input layer (i.e., sentence-matrix) is processed in a convolutional layer of multiple filters with different width, each of these results are fed into *Max Pooling* layers and finally the concatenation of them (previously flatten to be dimensional coherent) is projected into a dense layer. Our extension is to use a dilated filters in combination with normal ones, the intuition is that normal filter capture *adjacent words* features, while dilated one are able to capture relations between *non adjacent words*. This behaviour can't be achieved by the original Kim's model, because, even though the filters size can be changed, they will capture only features from adjacent words.
Regarding the architectural references in [1], the filter's number $|f|$ and their dimension (k, d), where k is the kernel size and d the dilation's unit, was optimized leading to the following results: $|f| = 5$, $f_1 = (2 \times 2, 0)$, $f_2 = (2 \times 2, 3)$, $f_3 = (3 \times 3, 1)$, $f_4 = (5 \times 5, 1)$, $f_5 = (7 \times 7, 1)$.

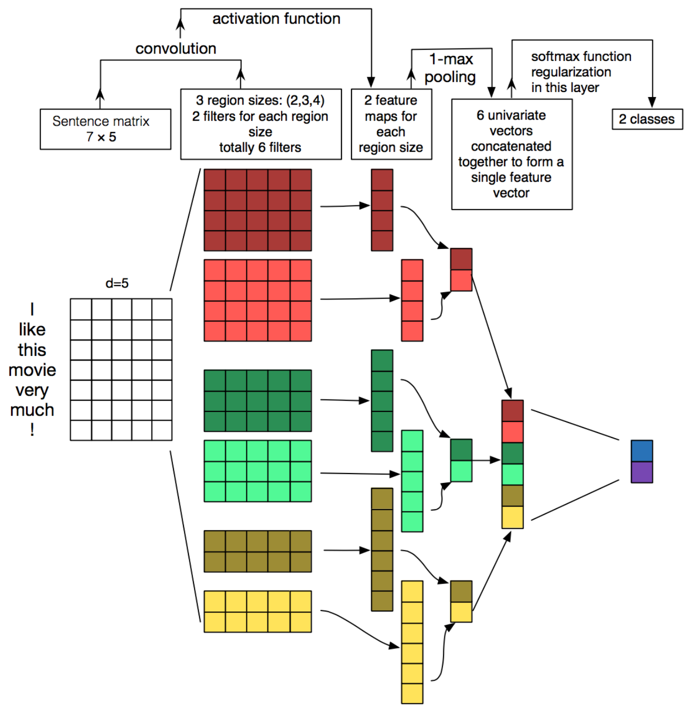

Fig. 1: [8] Illustration of a Convolutional Neural Network (CNN) architecture for sentence classification

Recurrent neural network. Long Short Term Memory (LSTM) and Bidirectional LSTM are types of Recurrent Neural Network (RNN) aiming at capture dynamic temporal behaviour. This behaviour suggest us to use them for the stance detection, in particular we use straightforward architectures made of an embedded input layer followed by an LSTM layer of 128 units, terminated by a dense layer for both normal and bidirectional models.

4 Evaluation

In this section we are going to illustrate the evaluation of developed systems regarding the modules design reported in section 3. First we illustrate the metric proposed by organizers for system's evaluation (Section 4.1), then we outline empirical results produced by a 10-fold cross validation over the given data set (Section 4.2), finally we report our performance at the shared task (Section 4.3).

4.1 Metrics

System evaluation metrics were given by the organizers and reported here in the following equations (1) to (6). Their choice was to use an $F_{1-macro}$ measure for stance detection, due to class unbalance, while a categorical accuracy for the gender detection.

$$Gender = accuracy = \frac{\sum TP + \sum TN}{\sum sample} \tag{1}$$

$$Stance = \frac{F_{1-macro}(Favor) + F_{1-macro}(Against)}{2} \tag{2}$$

$$F_{1-macro}(L) = \frac{1}{|L|} \sum_{l \in L} F_1(y_l, \hat{y}_l) \tag{3}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

$$precision = \frac{1}{|L|} \sum_{l \in L} Pr(y_l, \hat{y}_l) \tag{5}$$

$$recall = \frac{1}{|L|} \sum_{l \in L} R(y_l, \hat{y}_l) \tag{6}$$

where L is the set of classes, y_l is the set of correct label and $\hat{y}_l$ is the set of predicted labels.

4.2 Fine tuning process

Following, we describe the fine tuning process of our proposed model over possible combinations of pre-processing (Table 3), then we compare Kim's model against our extension (Table 4) and finally report the improvement over the use of a *data augmentation* technique (Table 5). For brevity of information only the evaluation of Dilated Kim's model over Spanish stance detection is reported, in details, the results are calculated from averaging three runs of a 10-fold cross validation over the complete data set. Nevertheless, the results obtained after the fine tuning process for all the models are reported in section 4.3, where their development performances are compared against the ones obtained in the *StanceCat* task.
Notations used in Table 3 refer to the one introduced in Section 3.1, where the listing of a notation means its use for the reported result. Regarding the tweet specific pre-processing, all the items have been substituted, with the exception for URL and RW that have been removed. We report the contribution of each analysed pre-processing alone.
From the analysis of Table 3 some relative observation can be made:

Table 3: Pre-processing fine tuning for the Dilated Kim's model from a three run of 10-fold cross validation over the development set. Results are in terms of average $F_{1-macro}$ score. The processing technique that brought a model's improvement has its result in bold.

Models	Pre-processing									
	Nothing	ST	SW	URL	RW	MT	HT	NUM	EM	SM
Dilated Kim	0.606	**0.615**	0.590	0.585	0.578	**0.610**	0.543	0.570	0.564	0.585

Table 4: Comparison of Kim's and Dilated Kim respect their best pre-processing tuning for stance&gender detection task. Results are averaged after three run of 10-fold cross validation over the development set in terms of averaged $F_{1-macro}$ score.

Models	Stance		Gender	
	ES	CA	ES	CA
Kim	0.624(±0.017)	0.630(±0.022)	0.634(±0.011)	0.655(±0.017)
Dilated Kim	0.658(±0.039)	0.659(±0.028)	0.652(±0.013)	0.715(±0.015)

- Most of common used preprocessing decrease model's performance, meaning that their information can be directly exploited by the model
- Only the stemming and mention remoning brough small improvements, therefore they will be used as a best tuning for our proposed model.

From the analysis of Table 4 we can outline a significant performance's improvement of our proposed model respect the original Kim's model.
Due to the fact that our development data set has few samples, to train our models we decided to apply a *data augmentation* technique that didn't rely over external data rather exploit the word embedding text representation. In details, we applied Gaussian noise to word embeddings and after every convolutional layers, and, to further improve performances, we take advantage of batch normalization. Results of this technique respect the Dilated Kim's model are reported in table 5.

Table 5: Data augmentation study for Dilated Kim's model over the Spanish stance detection development dataset. Results are averaged after three run of 10-fold cross validation over the development set in terms of averaged $F_{1-macro}$ score.

System	Nothing	Gaussian noise	Batch normalization
Dilated Kim	0.658 (± 0.039)	0.664 (± 0.043)	0.675 (± 0.049)

4.3 Competition results

For the system's submission, participants where allowed to send more than a model till a maximum of 5 possible runs, therefore in tables 6 and 7 we report our best performing systems (tuned following the process in section 4.2) for the StanceCat shared task.
Unfortunately, due to a submission error caught only after the official results were published, we didn't manage to be properly evaluated (the minus simbol in tables 6 and 7), therefore after the closing we asked organizers to evaluate some of our model to see how they would had performed (the test columns).

Table 6: Comparison of the best tuning model for the stance detection respect development and test set. The reported ranking refers to the absolute position over all submissions.

System	Development		Test			
	ES	CA	ES		CA	
			Score	Ranking	Score	Ranking
LSTM	0.443(±0.012)	0.489(±0.012)	-	-	-	-
Bi-LSTM	0.564(±0.035)	0.566(±0.035)	0.410	17/31	0.386	20/31
CNN	0.539(±0.030)	0.566(±0.030)	-	-	-	-
Kim	0.625(±0.019)	0.602(±0.019)	-	-	-	-
Dilated Kim	**0.675 (±0.049)**	**0.635 (±0.049)**	-	-	-	-

Table 7: Comparison of the best tuning model for the gender detection respect development and test set. The reported ranking refers to the absolute position over all submissions.

System	Development		Test			
	ES	CA	ES		CA	
			Score	Ranking	Score	Ranking
LSTM	0.579(±0.010)	0.648(±0.008)	-	-	-	-
Bi-LSTM	0.679(±0.025)	0.766(±0.028)	-	-	-	-
CNN	**0.756 (±0.027)**	**0.810 (±0.022)**	**0.736**	**1/21**	**0.457**	**4/17**
Kim	0.608(±0.017)	0.715(±0.014)	-	-	-	-
Dilated Kim	0.649(±0.029)	0.745(±0.039)	-	-	-	-

In table 6 (stance detection results), we see that our proposed Dilated Kim's model outperformed both recurrent and convolutional models, giving us insight for future developments. In table 7 (gender detection results), the best performance is achieved by a simple convolutional neural network, that from the test evaluation should had achieved the best result in the Spanish gender detection task.

5 Conclusions

In this paper we have presented our participation in the IberEval2017 Sta (StanceCat) shared task. Five distinct neural models were explored, in combination with different types of preprocessing. From the fine tuning process we derived that most of well know pre-processing technique are strongly model dependent, meaning that the preprocessing pipeline has to be optimized depending on the classifier. Our proposal of a dilation technique for NLP task, the Dilated Kim's model, seems to increase performances of CNN base classifiers.

References

1. Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).
2. Joulin, Armand, et al. "Bag of tricks for efficient text classification." arXiv preprint arXiv:1607.01759 (2016).
3. Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1 (ETMTNLP

'02), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI=http://dx.doi.org/10.3115/1118108.1118117
4. Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas. "Enriching Word Vectors with Subword Information" arXiv preprint arXiv:1607.04606 (2016).
5. Harris, Zellig S. "Distributional structure." Word 10.2-3 (1954): 146-162.
6. Bazzanella, Carla. "Oscillazioni di informalit e formalit: scritto, parlato e rete." Formale e informale. La variazione di registro nella comunicazione elettronica. Roma: Carocci (2011): 68-83.
7. Cerruti, Massimo, and Cristina Onesti. "Netspeak: a language variety? Some remarks from an Italian sociolinguistic perspective." Languages go web: Standard and non-standard languages on the Internet (2013): 23-39.
8. Zhang, Ye, and Byron Wallace. "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification." arXiv preprint arXiv:1510.03820 (2015).
9. C. Bosco, M. Lai, V. Patti, F. Rangel, P. Rosso (2016) Tweeting in the Debate about Catalan Elections. In: Proc. LREC workshop on Emotion and Sentiment Analysis Workshop (ESA), LREC-2016, Portoro, Slovenia, May 23-28, pp. 67-70.
10. F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, B. Stein (2016) Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: Balog K., Cappellato L., Ferro N., Macdonald C. (Eds.) CLEF 2016 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1609, pp. 750-784.
11. Mohammad, Saif M., Parinaz Sobhani, and Svetlana Kiritchenko. "Stance and sentiment in tweets." arXiv preprint arXiv:1605.01655 (2016).
12. Mohammad, Saif M., et al. "Semeval-2016 task 6: Detecting stance in tweets." Proceedings of SemEval 16 (2016).
13. Fisher Yu,Vladlen Koltun , Multi-Scale Context Aggregation by Dilated Convolutions