ELiRF-UPV at MultiStanceCat 2018

José-Ángel González[0000−0003−3812−5792], Lluís-Felip Hurtado[0000−0002−1877−0455], and Ferran Pla[0000−0003−4822−8808]

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

Abstract. This paper describes the participation of ELiRF-UPV team at the Spanish subtasks of the MultiModal Stance Detection in tweets on Catalan #1Oct Referendum workshop. Our best approach is based on Convolutional Neural Networks using word embeddings and polarity/emotion lexicons. We obtained competitive results on the Spanish subtask using only the text of the tweet, dispensing with contexts and images.

Keywords: Deep Learning · Stance Detection · Convolutional Neural Networks.

1 Introducction

Stance detection consists 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 Stance and Gender detection in Tweets on Catalan Independence (StanceCat-2017) [9].

MultiModal Stance Detection in tweets on Catalan #1Oct Referendum (MultiStanceCat) workshop is one of the tracks proposed at Ibereval 2018 workshop [10]. The aim of this track is to detect the stance with respect to the target "independence of Catalonia" in tweets written in Spanish and Catalan, furthermore, it is a multimodal task because both the text of the tweet and up to ten pictures of the user timeline could be taken into account for determining the stance.

2 Task Dataset

The corpus is composed by tweets labeled with respect to the stance of the Catalan first October Referendum (2017). There are three classes: AGAINST (AG), NEUTRAL (NE) and FAVOR (FA). These tweets are provided in Spanish and Catalan, however, we worked only with the Spanish subtask. Moreover, although context of the tweet and images are also provided by the organizers, we only used the text of the tweet.

From the official training corpus, we randomly selected 80% in order to train our models. The remaining 20% was used as development set. Table 1 shows the sample distribution per class in the Spanish corpus.

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

	Train	Dev
AG	1431	355
NE	758	213
FA	1360	320
Σ	3549	888

3 System Description

In this Section, we describe the two models used in the competition. Both models share the same preprocessing of the tweets by means of the TweetMotif [2] package. We applied a normalization step consisted on lowercasing the words, removing some language-specific characters such as accent, dieresis, special language characters, and normalizing Twitter-specific tokens (hashtags, user mentions and urls) by replacing them for a fixed word e.g. #1octL6 $\rightarrow$ #hashtag.

As first model for the experimentation, we used a Support Vector Machine (SVM) classifier with different representations of the tweets. Concretely, we used bag-of-word-ngrams and bag-of-char-ngrams with several values of n (including combination of ngrams e.g. bag-of-word-1-4grams means the concatenation of $n = [1, 4]$grams).

As second model for the experimentation, we used a Convolutional Neural Network (CNN) architecture inspired by the work presented in [12], with the aim of obtaining representations of the tweets similar to continuous versions of the bag-of-ngrams. We represented the tweets using Word2Vec distributed representations of words [3] [4]. Moreover, to enrich the system, we used several polarity/emotion lexicons combined with the word embeddings.

We used ELHPolar [8], ISOL [7], MLSenticon [1] and the Spanish version of NRC [6] as lexicons. As word embeddings, we trained a skip-gram model, with 300 dimensions for each word, from 87 million Spanish tweets collected for previous experimental work.

We represent each tweet x as a matrix $S \in \mathbb{R}^{n \times (d+v)}$, where n is the maximum number of words per tweet, d is the dimensionality of word embeddings and v is the dimensionality of the polarity/emotion features, that is, the number of polarity/emotion lexicons. In order to obtain this representation, we use an embedding model $h(w) \in \mathbb{R}^d$ and a set of lexicons $h^{'}(w) = [h^{'}_1(w), h^{'}_2(w), ..., h^{'}_l(w)] \in \mathbb{R}^v$, where $h^{'}_k(w)$ is the polarity value of the word w in the lexicon k.

Therefore, given a tweet x with n tokens, $x = w_1, w_2, ..., w_n$, we represent it as a matrix S in which, each row i is the concatenation of the embedding of w_i ($h(w_i)$) and a vector with the polarity values of w_i in each lexicon ($h^{'}(w_i)$),

$S = [h(w_1)|h^{'}(w_1), h(w_2)|h^{'}(w_2), ..., h(w_n)|h^{'}(w_n)]$. In the case where a word w_i is out of vocabulary for the embedding models, we replace its embedding by the embedding of the word "unknown", $h(w_i) = h(\text{“}unknown\text{”})$. Similarly, if w_i is not included in any lexicon, $h^{'}(w_i) = [0, 0, ..., 0] \in \mathbb{R}^v$.

Due to the variable length of the tweets, we used zero padding at the start of a tweet if it does not reach the maximum specified length. Otherwise, if the length of a tweet is greater than the maximum, we only consider the first n words of the tweet. In this task, the average number of words per tweet is $n_{avg} = 18.5$, and the maximum length is $n_{max} = 34$. We decided to set the length $n = 26$ which is the mean of n_{avg} and n_{max}.

Regarding the CNN architecture, we applied one-dimensional convolutions with variable height filters in order to extract the temporal structure of the tweet over several region sizes. Figure 1 summarizes the model architecture and its hyperparameters.

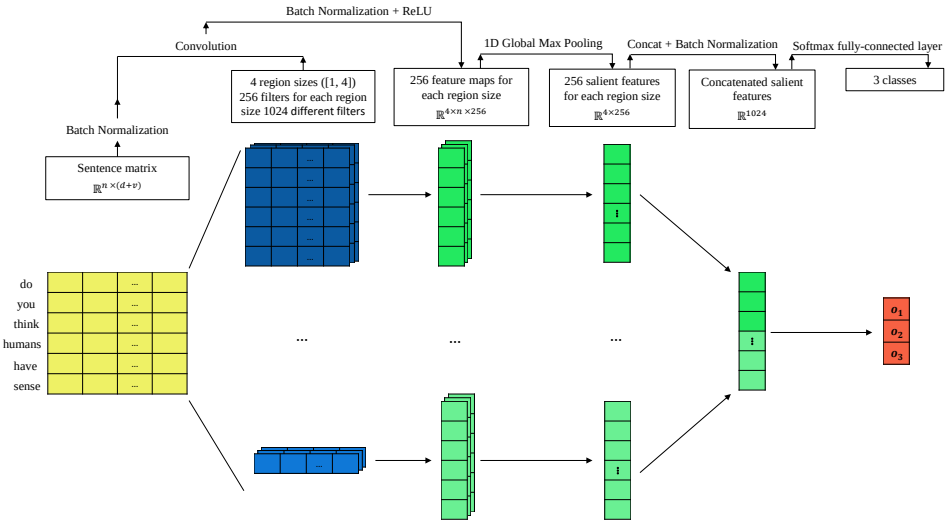

Fig. 1. CNN architecture for multi-label classification.

As can be seen in Figure 1, we used 4 different region sizes (the filter height range from 1 to 4) and 256 filters for each region size. We used this range of region sizes because, in the development phase, the best baseline was SVM using bag-of-word-1-4grams. After the filters were applied, we obtained 256 output feature maps for each region size.

In order to extract the most salient features for each region size, we applied 1D Global Max Pooling to the feature maps of each region size. Therefore, we obtained 4 vectors with 256 components, that were concatenated and used as

input to a fully-connected layer which performs the classification task. We used a softmax activation function to model the posterior distribution of each class at the output layer.

4 Experimental results

In this section, we describe the experimental work conducted by ELiRF team in the MultiStanceCat task. In addition, we present a study of the performance of our best system in the competition.

Table 2 summarizes the results obtained in the development phase. Three different classifiers were considered: Linear SVM with bag-of-ngrams of words, Linear SVM with bag-of-ngrams of chars and CNNs.

For the SVM approaches we tested different values of n. With respect to the CNN we explored three loss functions: Cross Entropy (CCE), Mean Squared Error (MSE) and differentiable approximation of the F_1 measure (SMF_1).

Table 2. Results obtained with the different approaches considered in the development phase.

Experiments			F_1 (AG)	F_1 (NE)	F_1 (FA)	$\frac{F_1(AG)+F_1(FA)}{2}$
Linear SVM	**Word Ngrams**	**1-1-grams**	59.48	46.19	57.06	58.27
		1-2-grams	62.95	49.73	56.07	59.51
		1-3-grams	64.09	52.25	59.80	61.94
		1-4-grams	64.87	49.86	59.30	**62.09**
	Char Ngrams	**1-1-grams**	47.46	39.90	53.32	50.39
		1-2-grams	57.25	48.49	37.40	47.32
		1-3-grams	55.91	45.54	51.52	53.72
		1-4-grams	58.63	48.02	51.52	53.72
		1-5-grams	59.55	48.76	55.59	57.57
		1-6-grams	60.76	49.12	57.36	**59.06**
CNN	**Embeddings + Lexicons**	**MSE**	65.27	50.91	60.56	62.91
		CCE	64.37	50.56	61.12	62.75
		SMF_1	67.13	50.15	63.51	**65.32**

It can be observed in Table 2 that generally bag-of-chars performs worse than bag-of-words. Note that CNN models outperform the results achieved by the SVM classifiers. Moreover, CNN classifier with SMF_1 loss function outperforms the results of all the other classifiers. However, a deeper study about which factors such as embeddings or lexicons are more relevant in the results would be interesting.

It can be also observed that the value of the F_1 measure for the NEUTRAL class ($F_1(NE)$) is generally lower than the F_1 measures for AGAINST and FAVOR classes ($F_1(AG), F_1(FA)$). We hypothesize this is due to the fact that NEUTRAL class has less samples in the corpus. However, low values of $F_1(NE)$ measure do not affect the official evaluation measure that is defined as the average between $F_1(AG)$ and $F_1(FA)$.

For the Spanish subtask competition, we selected the best CNN and SVM models according to the results obtained in the development phase. Concretely, our first run (ELiRF-1) was the CNN model trained using SMF_1 loss function. As a second run (ELiRF-2) we selected a Linear SVM with bag-of-word-1-4grams.

Table 3 shows the confusion matrices of the two submitted systems. It can be observed that both systems confuse the NEUTRAL and the AGAINST classes in a similar way. The best performance achieved by ELiRF-1 run is because it predicts better the FAVOR class.

Table 3. Confusion matrices for ELiRF-1 and ELiRF-2 systems.

		ELiRF-1			ELiRF-2		
		Predicted					
		AG	NE	FA	AG	NE	FA
Truth	AG	240	18	97	241	21	93
	NE	54	86	73	56	86	71
	FA	66	26	228	91	25	204

We have also performed a study of the samples that ELiRF-1 system misclassified with high confidence. Some of these samples are shown in Table 4. We think that in some cases, errors could be avoided by considering hashtags (sample 5, #4gatos) or user mentions (error 2, @CatalunyaPlural). Unfortunately, we have not included this information in our models.

Table 4. Examples of misclassifications with maximum confidence.

(1) Class FA, Predicted AG with 100% confidence: #1octL6 No puede haber referéndum pactado por políticos Estáis vendiendo una falacia La soberanía es nuestra

(2) Class FA, Predicted NE with 99.66% confidence: La policía nacional cita a declarar al líder del Partido Pirata de Catalunya por el referéndum https://t.co/dOq4igEDKf @CatalunyaPlural #1O

(3) Class NE, Predicted FA with 99.99% confidence: Gran vídeo para aquellos que sois normales. #hispanoMola #A3dias1OctARV #1octubreARV https://t.co/e8618WfUmA

(4) Class NE, Predicted AG with 99.99% confidence: #1octL6 yo pensaba que en la jornada de reflexión no se puede debatir ????????

(5) Class AG, Predicted FA with 100% confidence: @InesArrimadas @carrizosacarlos ya era hora que vuestra mayoría silenciosa saliera a la calle. #4gatos #1O... https://t.co/RdRdn8eWnH

(6) Class AG, Predicted NE with 99.99% confidence: De la Revolución de las Sonrisas al Conflicto Civil https://t.co/HwVbPBmEKM #1octL6 https://t.co/URtNC6YOXv

Table 5 shows the results on the test set for all the participating teams in the Spanish task. The ELiRF-1 run obtained competitive results without using the text of previous and next tweets or the images in the user timeline. Moreover, we can observe that the context seems to be useful for this task because all the best participating teams used this information. Finally, we would like to highlight the great difference observed in the results obtained on the development and the test sets. We have no explanation for this, but we think that a study about this aspect should be done when the test set will be available.

Table 5. Test results on the Spanish subtask.

Team	**Run**	**Macro** F_1
uc3m	text+context	28.02
CriCa	context	27.15
Casacufans	text+context+images	27.09
Casacufans	text+context	26.98
ELiRF-1	text	22.74
uc3m	text	22.47
CriCa	text	22.06
Casacufans	text	21.94
ELiRF-2	text	21.32

5 Conclusions and Future Work

In this paper, we have presented the participation of the ELiRF team at Multi-StanceCat track of the IberEval workshop. Our team participated in the Spanish subtask of this track and competitive results were achieved using only the text of the tweets. Our best approach is based on CNN with sequential representation of the tweets using word embedding, and polarity/emotion lexicons.

As future work, we plan to include the context of the tweet in our deep learning system in a similar way as Hierarchical Attention Networks [11] do. Moreover, we think that data augmentation could help to improve the performance of the models.

We have observed that hashtags and user mentions contains relevant information for this task. For this reason, as future work, we want to explore the inclusion of this information in the tweet representation.

6 Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC (TIN2017-85854-C4-2-R). Work of José-Ángel González is also financed by Universitat Politècnica de València under grant PAID-01-17.

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)
2. Krieger, M., Ahn, D.: Tweetmotif: exploratory search and topic summarization for twitter. In: In Proc. of AAAI Conference on Weblogs and Social (2010)
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., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 31–41 (2016)
6. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence **29**(3), 436–465 (2013)
7. Molina-González, M.D., Martínez-Cámara, 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)
8. Saralegi, X., San Vicente, I.: Elhuyar at tass 2013. In: XXIX Congreso de la Sociedad Espaola de Procesamiento de lenguaje natural, Workshop on Sentiment Analysis at SEPLN (TASS2013). pp. 143–150 (2013)
9. Taulé, M., Martí, M.A., Rangel, F.M., Rosso, P., Bosco, C., Patti, V., et al.: Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In: 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017. vol. 1881, pp. 157–177. CEUR-WS (2017)
10. Taulé, M., Rangel, F.M., Martí, M.A., Rosso, P.: Overview of the task on multimodal stance detection in tweets on catalan #1oct referendum'. In: Third Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018 (2018)
11. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
12. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 253–263. Asian Federation of Natural Language Processing (2017), http://aclweb.org/anthology/I17-1026