ELiRF-UPV at IberEval 2017: Classification Of Spanish Election Tweets (COSET)

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 Classification Of Spanish Election Tweets (COSET) task. We tested several approaches based on different classifiers and features representations. Our main approach is based on neural networks, concretely, Multilayer Perceptrons (MLP) with bag-of-words representation of the tweets. Our system achieved the best score on the test set of the COSET task with 64.82 of macro-F_1.

Keywords: Neural Networks, MLP, bag-of-words

1 Introduction

Text Categorization (TC) is a well-known and widely applied machine-learning technique for classifying textual documents into categories [12]. The goal of TC is the classification of documents into a fixed number of predefined categories.

Recently, there have been several works that adapt traditional TC methods for classifying Twitter posts into a predefined set of generic classes [3] [8].

The aim of COSET shared task at Ibereval 2017 [4] is to classify a corpus of political tweets in five categories: political issues related to electoral confrontation, sectoral policies, personal questions about the candidates, issues of the electoral campaign and other issues. This data has been collected during the Spanish electoral campaign of 2015.

The present paper describes the participation of the ELiRF-UPV team at COSET shared task, the way in which the task has been addressed and the results obtained using different approaches.

2 Corpus Description

The corpus is composed by tweets belonging to five categories: *political issues*, related to the most abstract electoral confrontation; *policy issues*, about sectorial policies; *personal issues*, on the life and activities of the candidates; *campaign issues*, related with the evolution of the campaign; and *other issues*.

These tweets are written in Spanish and were collected during the general Spanish elections of 2015. The organization provided this corpus divided in three

parts. The training set (2242 tweets), the development set (250 tweets) and the test set (624 tweets).

Table 1. Topic distribution of the tweets in the Training and Development sets.

Topic	#tweets Training	#tweets Development
Political issues	530 (23%)	57 (23%)
Policy issues	786 (35%)	89 (35%)
Campaign issues	511 (23%)	71 (28%)
Personal issues	152 (7%)	9 (4%)
Other issues	263 (12%)	25 (10%)
Total	2242 (100%)	250 (100%)

Note that the topic distribution of the corpus is not uniform (see Table 1). The three majority classes (*political issues*, *policy issues* and *personal issues*) represent more than the 80% of the samples. Moreover, both training and development sets have similar distributions of tweets per category.

3 System Description

In this section we describe the main characteristics of the system developed to the COSET task competition. This description includes the preprocessing used, the feature selection process and the different models that were taken into account during the tuning phase.

3.1 Preprocessing

In the preprocessing conducted all the tweets were converted to lowercase and the accents were removed. In all cases, Url's, user's mentions, numbers, exclamations and interrogations are replaced by a specific label. On the contrary, hashtags were not replaced because we observed that hashtags were relevant in the classification process.

3.2 Models and Feature Selection

Different classifiers and tweets representations were considered. Three models were tested: Support Vector Machines (SVM) with linear kernel, Multilayer Perceptron (MLP), and deep learning models (CNN+LSTM), that we have recently used in Sentiment Analysis tasks [5].

We selected the most appropriate tweets representations for each considered model. This way, we tested SVM and MLP models with bag-of-ngrams ($n \geq 1$) of words and characters and TF-IDF weighting; and CNN+LSTM with sequences of Spanish Wikipedia embeddings [9] [10] [13] [11] and sequences of one-hot vectors.

Table 2 shows the results of the different combinations considered in the development phase.

Table 2. Results on the development set with different models and representations.

System	Features	Macro-F_1
MLP	**BOW**	**68.23**
MLP	TF-IDF	59.31
MLP	Collapse embeddings	46.73
SVM-Linear	BOW	58.66
SVM-Linear	TF-IDF	55.16
SVM-Linear	Collapse embeddings	48.85
CNN+LSTM	One-Hot sequence	40.68
CNN+LSTM	Embedding sequence	55.43

The best results were obtained using Multilayer Perceptron (MLP) models with bag-of-words (BOW) as representation of tweets. Support Vector Machines also presented good results, in particular by using BOW representations. Contrary to what we expected, deep learning models obtained the worst results.

Specifically, our best model was a Multilayer Perceptron of 3 hidden layers with 128 neurons and ReLU as activation function in all layers (except Softmax in the output layer).

The optimization method was Adagrad [2], which has provided the best results in similar TC tasks, and categorical cross-entropy as a loss function. In addition, a scaling of the loss function is used to deal with the problem of unbalanced classes. This scaling method has the following form: $f_{loss}(x) \cdot \log(\mu \cdot \frac{n_r}{n_c})$ where n_r is the number of samples of the majority class and n_c is the number of samples of the class of sample x.

4 Results

In view of the results obtained during the tuning phase, we decided to send four systems with the same base architecture (Multilayer Perceptron with bag-of-words representation), but with some minor changes:

- **System-1**. A MLP trained using all data (training and development sets) with the parameters adjusted on the development set.
- **System-2**. A MLP trained using only the training set.
- **System-3**. A MLP trained using with all data, but without scaling the loss function.
- **System-4**. A majority voting schema of the three previous systems.

The official results achieved by our four systems are shown in Table 3. We have also included, in parenthesis, the position of each system in the ranking of the COSET competition.

Table 3. Results of ELiRF-UPV systems.

System	Macro-F_1
System-1	**64.82 (1)**
System-2	62.33 (5)
System-3	63.33 (4)
System-4	64.00 (2)

The best results of the competition were obtained by our System-1, which achieved 64.82 of macro-F_1 on the test set. The scaling of the loss function adjusted on the development set also behaved well when the system is trained with all samples.

System-2 obtained the worst results of our submissions (5th position over 39 submissions). Note that, this was the system learned with less data. System-3 was learned with the same samples of System-1 but without scaling the loss function. From the results achieved by System-3, we want to highlight the importance of dealing with the problem of unbalanced classes.

The majority voting proposed (System-4) could not outperform the results of the best model (System-1). Perhaps, a more sophisticated combination method could obtain better results.

Finally, our team has participated in the second evaluation phase proposed by the COSET organization. In this second phase, a new corpus was automatically labeled with the agreement of four of the five participant runs (80% of agreement). A 65.91% of the considered tweets achieved the agreement criterion. Therefore, the final corpus size was 10,417,058 tweets. Table 4 shows the results of the second evaluation phase, sorted by macro-F_1 measure. We achieved the best two results in terms of macro-F_1 in this second phase. In this case, the best system, ELiRF-UPV.1, was trained with the whole COSET corpus and the second system, ELiRF-UPV.2, was learned using only the training set of the COSET corpus.

Table 4. Results of ELiRF-UPV systems with large corpus.

Team	macro-F_1
ELiRF-UPV.1	**95.86**
ELiRF-UPV.2	95.23
Team 17	94.82
atoppe	89.60
LTRC_IIITH	85.09

5 Conclusions and Future Work

We have presented the participation of ELiRF-UPV team at Classification Of Spanish Election Tweets (COSET) task. We tested several approaches. The best results were obtained by Multilayer Perceptrons models with bag-of-words representation. Our system achieved the best score on the competition with 64.82 of macro-F_1.

In this task, the scaling of the loss function seems to be a key factor to improve the results, as it can be observed in the final ranking, where the system that uses this scaling performs better than the systems that do not use it.

As future work, we plan to study the generation of synthetic samples to tackle with the unbalance problem. We want to test data-augmentation techniques and generative models as SMOTE [1], GAN [6], or VAE [7].

Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under project ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics, TIN2014-54288-C4-3-R.

References

1. Bowyer, K.W., Chawla, N.V., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. CoRR abs/1106.1813 (2011), http://arxiv.org/abs/1106.1813
2. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12, 2121–2159 (Jul 2011), http://dl.acm.org/citation.cfm?id=1953048.2021068
3. Fiaidhi, J., Mohammed, S., Islam, A., Fong, S., Kim, T.h.: Developing a hierarchical multi-label classifier for twitter trending topics. International Journal of U-& E-Service, Science & Technology 6(3) (2013)
4. Giménez, M., Baviera, T., Llorca, G., Gámir, J., Calvo, D., Rosso, P., Rangel, F.: Overview of the 1st Classification of Spanish Election Tweets Task 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, Murcia, Spain (2017)
5. 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)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. CoRR abs/1406.2661 (2014), http://arxiv.org/abs/1406.2661
7. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR abs/1312.6114 (2013), http://arxiv.org/abs/1312.6114

8. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.: Twitter trending topic classification. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 251–258. ICDMW '11, IEEE Computer Society, Washington, DC, USA (2011), http://dx.doi.org/10.1109/ICDMW.2011.171
9. 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
10. 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
11. 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
12. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (Mar 2002), http://doi.acm.org/10.1145/505282.505283
13. Wikipedia: Wikipedia spanish dumps (2017), https://dumps.wikimedia.org/eswiki/, [Online; accessed 18-May-2017]