ELiRF-UPV at TASS 2019: Transformer Encoders for Twitter Sentiment Analysis in Spanish

José-Ángel González, Lluís-Felip Hurtado, and Ferran Pla

VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València
{jogonba2,lhurtado,fpla}@dsic.upv.es

Abstract. This paper describes the participation of the ELiRF research group of the Universitat Politècnica de València in the TASS 2019 Workshop, framed within the XXXV edition of the International Congress of the Spanish Society for the Processing of Natural Language. We present the approach used for the Monolingual InterTASS task of the workshop, as well as the results obtained and a discussion of them. Our participation has focused mainly on employing the encoders of the Transformer model, based on self-attention mechanisms, achieving competitive results in the task addressed.

Keywords: Twitter · Sentiment Analysis · Transformer Encoders.

1 Introduction

Sentiment Analysis workshop at SEPLN (TASS) has been proposing a set of tasks related to Twitter sentiment analysis in order to evaluate different approaches presented by the participants. In addition, it develops free resources, such as, corpora annotated with polarity, thematic, political tendency or aspects, which are very useful for the comparison of different approaches to the proposed tasks.

In this eighth edition of the TASS [3], several tasks are proposed for global sentiment analysis about different Spanish variants. The organizers propose two different tasks: 1) Monolingual sentiment analysis and 2) crosslingual sentiment analysis. In this way, in the first task, only a specific language can be used to train and to evaluate the system; in contrast, in the second task, any combination of the corpus can be used to train the systems. Thus, for both tasks, the organizers provide five different corpus of tweets written in Spanish variants from Spain, Costa Rica, Peru, Uruguay and Mexico.

This article summarizes the participation of the ELiRF-UPV team of the Universitat Politècnica de València only for the first task. Our approach uses

state-of-the-art approaches that has provided competitive results in English sentiment analysis and machine translation [8] [1].

The rest of the article is structured as follows. Section 2 presents a description of the addressed task. In section 3 we describe the architecture of the proposed system. Section 4 summarizes the conducted experimental evaluation and the achieved results. Finally, some conclusions and possible future works are shown in section 5.

2 Task description

The organization has defined two subtasks: Task 1, monolingual SA, and Task 2, crosslingual SA. These tasks consists of assigning a global polarity to tweets (**N**, **NEU**, **NONE** and **P**). In Task 1 only one Spanish variant can be used, both for training and testing the system. In contrast, in Task 2 any combination of Spanish variants can be considered with the only restriction that those considered in the training set can not be used in the test set.

For both subtasks, five different corpora were considered for several Spanish variants. First, the InterTASS-ES corpus (Spain) composed of a training partition of 1125 samples, a validation of 581 samples and a test set consisting of 1706 samples. InterTASS-CR (Costa Rica) composed of 777 training samples, 390 for validation and 1166 for testing. InterTASS-PE (Peru), formed by 966 samples of training, 498 of validation and 1464 of test. InterTASS-UY (Uruguay), which contains 943 training samples, 486 validation and 1428 tests. Finally, InterTASS-MX (Mexico), with 989 training samples, 510 validation and 1500 test samples.

The tweet distribution according to their polarity in the InterTASS corpus training sets is shown in Table 1.

Table 1: Distribution of tweets in the training sets of InterTASS for all the Spanish variants.

	ES	CR	PE	UY	MX
N	474	310	228	367	505
NEU	140	91	170	192	79
NONE	157	155	352	94	93
P	354	221	216	290	312
Σ	1125	777	966	943	989

As can be seen in Table 1, the training corpora are unbalanced and they have a bias to the N and P classes, except in the InterTASS-PE corpus, where the most frequent class is $NONE$. Moreover, the class NEU is always the least populated except in the case of Uruguay.

3 System

In this section, we discuss the system architecture proposed to address the first task of TASS 2019 as well as the description of the resources used and the preprocessing applied to the tweets.

3.1 Resources and preprocessing

In order to learn a word embedding model from Spanish tweets, we downloaded 87 million tweets of several Spanish variants. To provide the embedding layer of our system with a rich semantic representation on the Twitter domain, we use 300-dimensional word embeddings extracted from a skip-gram model [9] trained with the 87 million tweets by using Word2Vec framework [4].

3.2 Transformer Encoders

Our system is based on the Transformer [11] model. Initially proposed for machine translation, the Transformer model dispenses with convolution and recurrences to learn long-range relationships. Instead of this kind of mechanisms, it relies on multi-head self-attention, where multiple attentions among the terms of a sequence are computed in parallel to take into account different relationships among them.

Concretely, we use only the encoder part in order to extract vector representations that are useful to perform sentiment analysis. We denote this encoding part of the Transformer model as Transformer Encoder. Figure 1 shows a representation of the proposed architecture for sentiment analysis.

The input of the model is a tweet $X = \{x_1, x_2, ..., x_T \ : x_i \in \{0, ..., V\}\}$ where T is the maximum length of the tweet and V is the vocabulary size. This tweet is sent to a d-dimensional fixed embedding layer, E, initialized with the weights of our embedding model. Moreover, to take into account positional information we also experimented with the sine and cosine functions proposed in [11]. After the combination of the word embeddings with the positional information, dropout [10] was used to drop input words with a certain probability p. On top of these representations, Nx transformer encoders are applied, which relies on multi-head scaled dot-product attention with h different heads. To do this we used an architecture similar to the one described in [11]. It includes the layer normalization [2] and the residual connections.

Due to a vector representation is required to train classifiers on top of these encoders, a global average pooling mechanism was applied to the output of the last encoder, and it is used as input to a feed-forward neural network, with only one hidden layer, whose output layer computes a probability distribution over the the four classes of the task $\mathbb{C} = \{P, N, NEU, NONE\}$.

We use Adam as update rule with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and Noam [11] as learning rate schedule with 5 *warmup_steps*. The weighted cross entropy is used as loss function. Only the class distribution of the Spanish variant is considered to weight the cross entropy that is used for all language variants.

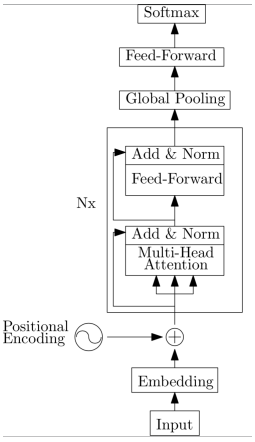

Fig. 1: The Transformer Encoder system for TASS 2019.

4 Experiments

We fixed some hyper-parameters to carry out the experimentation, concretely: $batch\_size = 32$, $d_k = 64$, $d_{ff} = d$ and $T = 50$. Another hyper-parameters such as p or $warmup\_steps$ were set following some results obtained in preliminary experiments to $p = 0.7$, $warmup\_steps = 5$ epochs and $h = 8$.

Moreover, we compair our proposal, which is based on transformer encoders (TE), with another deep learning systems such as Deep Averaging Networks (DAN) [7] and Attention Long Short Term Memory Networks [6] (Att-LSTM) that are commonly used in related text classification tasks obtaining very competitive results. Concretely, these implementations are the systems proposed by our team in the TASS2018 edition, which achieved very competitive results [5].

In order to study how some system mechanisms (positional encodings) or hyper-parameters (Nx) affect the results obtained in terms of macro-F_1 (MF_1), macro-recall (MR), macro-precision (MP) and Accuracy (Acc) we conducted some additional experimentation. Concretely, we removed the positional information and we used $Nx \in \{1, 2\}$ encoders. All the configurations were applied only to the Spanish subtask and the best two configurations are used also in the remaining subtasks. All these results are shown in Table 2.

As it can be seen in Table 2 for systems 1-TE-Pos and 2-TE-Pos on subtask ES, the use of positional information decreases the system performance. This seems to indicate that the positional information, represented by sine and cosine functions added to the word embeddings, is useless to the classifier. However, the results obtained by Att-LSTM, which takes into account the positional information by its internal memory, obtains better results than the 1-TE-Pos and 2-TE-Pos in almost all the metrics. These results show that the way the positional information is considered affects the performance of the systems in this task.

Table 2: Results of the experimentation in the different variants.

	MP	MR	MF_1	Acc
		ES		
DAN	47.66	48.46	47.94	56.28
Att-LSTM	50.00	48.14	48.83	58.00
1-TE-NoPos	52.80	**54.38**	**53.34**	60.75
1-TE-Pos	46.26	46.56	46.25	55.94
2-TE-NoPos	**52.85**	53.03	51.47	**61.27**
2-TE-Pos	47.31	48.79	47.71	56.11
		PE		
1-TE-NoPos	**49.06**	**50.43**	**49.51**	**54.62**
2-TE-NoPos	46.29	46.00	44.92	46.79
		CR		
1-TE-NoPos	**55.36**	**56.10**	**54.56**	**58.46**
2-TE-NoPos	52.14	52.36	51.71	55.13
		UY		
1-TE-NoPos	54.71	**56.63**	**54.83**	57.20
2-TE-NoPos	**55.82**	53.56	54.29	**58.64**
		MX		
1-TE-NoPos	**53.59**	55.03	**54.10**	**63.52**
2-TE-NoPos	52.78	**57.34**	54.07	60.78

The best results in terms of MR are achieved by the 1-TE-NoPos model. Due to this fact, the 1-TE-NoPos model outperforms 2-TE-NoPos model also in terms of MF_1, although the 2-TE-NoPos model achieves better results in the MP measure. This behavior is observed in almost all the Spanish variants, except on the MX subtask, where both models obtain similar results in terms of MF_1.

Moreover, in the ES variant, several configurations of the TE model outperforms the systems proposed by our team in previous editions of TASS (DAN and Att-LSTM) by a margin of $\sim$5 points of MF_1, mainly due to the improvement ($\sim$6 points) in terms of MR and MP (improvement of $\sim$3 points).

In Table 3, the results at class level for each variant, obtained with our best model (1-TE-NoPos), are shown. It is interesting to observe the improvements achieved by our system for the class $NONE$ compared to our results in previous editions for this class. Generally, the results for the class N are better than those obtained on the other classes, except in the PE variant. In this case the $NONE$ class is the easiest to detect due to this class is most frequent in the corpus. The results for P class are generally better than those for classes NEU and $NONE$, except on the PE variant. As it is observed in all the previous editions of TASS[5], the NEU class obtains the worse results.

The confusion matrix of our best system (1-TE-NoPos) for the ES variant is shown in Table 4. It is possible to see that the worse classified class (NEU) is usually confused with the classes N and P. This seems to indicate that our

Table 3: Results at class level, for each sub-task, of the best model 1-TE-NoPos (0 refers to N class, 1 to NEU, 2 to NONE and 3 to P).

	P_0	P_1	P_2	P_3	R_0	R_1	R_2	R_3	F_{1_0}	F_{1_1}	F_{1_2}	F_{1_3}
ES	**73.03**	30.56	46.34	61.25	**73.31**	26.51	59.38	58.33	**73.17**	28.39	52.05	59.76
PE	51.40	27.27	**64.88**	52.67	51.40	26.79	57.83	**65.71**	51.40	27.03	**61.15**	58.47
CR	**74.58**	27.87	46.09	72.92	61.54	30.91	**73.61**	58.33	**67.43**	29.31	56.68	64.81
UY	**69.70**	34.51	50.00	64.64	47.92	43.33	58.85	**76.47**	56.79	38.42	54.05	**70.06**
MX	**73.93**	30.91	44.07	65.47	**75.40**	33.33	54.17	57.23	**74.66**	32.08	48.60	61.07

model detects the presence of sentiment (positive or negative), but is unable to detect when both classes are neutralized.

Table 4: Confusion matrix for 1-TE-NoPos model on the ES development set.

	N	**NEU**	**NONE**	**P**
N	195	25	18	28
NEU	25	22	13	23
NONE	9	6	38	11
P	38	19	13	98

Finally, the system 1-TE-NoPos was used for labeling the test set of each variant. The results obtained by this model (MF_1, MP, and MR) and the ranking of our system in the competition are shown in Table 5. As it can be seen, our system is ranked as first for the ES subtask and second in all the remaining variants.

Table 5: Results and ranking of our system on the test sets.

	$\boldsymbol{MF_1}$	**MP**	**MR**	**Rank**
ES	50.70	50.50	50.80	1/9
CR	49.60	49.80	49.30	2/9
PE	44.70	45.60	43.90	2/9
UY	51.50	49.70	53.60	2/7
MX	50.10	49.00	51.20	N/A

5 Conclusions

We have proposed a system based on the encoder part of the Transformer architecture in order to extract useful word representations that are discriminative to perform sentiment analysis on tweets from several Spanish variants. The results

obtained by our system are very promising, being the first or second ranked system on almost all the Spanish variants. This is especially significant, considering that these results have been obtained without an extensive experimentation on the hyperparameters of the model and these hyperparameters were only tuned on the ES subtask. This opens the door to future improvements by exploring modifications on the architecture and its hyperparameters.

Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC (TIN2017-85854-C4-2-R) and by the GiSPRO project (PROMETEU/2018/176). Work of José-Ángel González is financed by Universitat Politècnica de València under grant PAID-01-17.

References

1. Ambartsoumian, A., Popowich, F.: Self-attention: A better building block for sentiment analysis neural network classifiers. In: WASSA@EMNLP (2018)
2. Ba, L.J., Kiros, R., Hinton, G.E.: Layer normalization. CoRR **abs/1607.06450** (2016), http://arxiv.org/abs/1607.06450
3. Díaz-Galiano, M.C., et al.: Overview of tass 2019. CEUR-WS, Bilbao, Spain (2019)
4. González, J., Hurtado, L., Pla, F.: ELiRF-UPV en TASS 2017: Análisis de Sentimientos en Twitter basado en Aprendizaje Profundo (ELiRF-UPV at TASS 2017: Sentiment Analysis in Twitter based on Deep Learning). In: Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2017, co-located with 33nd SEPLN Conference (SEPLN 2017), Murcia, Spain, September 18th, 2017. pp. 29–34 (2017), http://ceur-ws.org/Vol-1896/p2_elirf_tass2017.pdf
5. González, J., Hurtado, L., Pla, F.: ELiRF-UPV en TASS 2018: Análisis de Sentimientos en Twitter basado en Aprendizaje Profundo (ELiRF-UPV at TASS 2018: Sentiment Analysis in Twitter based on Deep Learning). In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018. pp. 37–44 (2018), http://ceur-ws.org/Vol-2172/p2_elirf_tass2018.pdf
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, http://dx.doi.org/10.1162/neco.1997.9.8.1735
7. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1681–1691. Association for Computational Linguistics, Beijing, China (Jul 2015). https://doi.org/10.3115/v1/P15-1162, https://www.aclweb.org/anthology/P15-1162
8. Letarte, G., Paradis, F., Giguère, P., Laviolette, F.: Importance of self-attention for sentiment analysis. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 267–275. Association for Computational Linguistics, Brussels, Belgium (Nov 2018), https://www.aclweb.org/anthology/W18-5429

9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013), http://dl.acm.org/citation.cfm?id=2999792.2999959
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)