ELiRF-UPV at TASS 2020: TWilBERT for Sentiment Analysis and Emotion Detection in Spanish Tweets

José-Ángel González[a], José Arias Moncho, Lluís-Felip Hurtado[a] and Ferran Pla[a]

[a] *VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València*

Abstract

This paper describes the participation of the ELiRF research group of the Universitat Politècnica de València in the TASS 2020 Workshop, framed within the XXXVI edition of the International Conference of the Spanish Society for the Processing of Natural Language (SEPLN). We present the approach used for the Monolingual Sentiment Analysis and Emotion Detection tasks of the workshop, as well as the results obtained. Our participation has focused mainly on employing an adaptation of BERT for text classification on the Twitter domain and the Spanish language. This system, that we have called TWilBERT, shown systematic improvements of the state of the art in almost all the tasks framed in the SEPLN conference of previous years, and also obtains the most competitive performance in the tasks addressed in this work.

Keywords

Twitter, Sentiment Analysis, Emotion Detection, TWilBERT,

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 ninth edition of the TASS, two different tasks are proposed both for Sentiment Analysis in several Spanish variants [1] and Emotion Detection.

This article summarizes the participation of the ELiRF-UPV team of the Universitat Politècnica de València. Following the competitive performance obtained the past edition by using non-pretrained Transformer Encoders [1], we decided to extend our approach by pre-training large Transformer Encoders in a similar way as BERT model. Thus, our approach (TWilBERT) is based on the fine-tuning of a pre-trained adaptation of the BERT model for Twitter and the Spanish language.

The rest of the article is structured as follows. Section 2 presents a description of the addressed tasks. In section 3 we describe our proposal (TWilBERT) and the baseline system we used

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: jogonba2@dsic.upv.es (J. González); joarmon1@inf.upv.es (J.A. Moncho); lhurtado@dsic.upv.es (L. Hurtado); fpla@dsic.upv.es (F. Pla)

ORCID: 0000-0003-3812-5792 (J. González); 0000-0002-1877-0455 (L. Hurtado); 0000-0003-4822-8808 (F. Pla)

[1] In this work, we only considered the mono-lingual version of this task

Table 1
Distribution of tweets in the training sets of the first task for all the Spanish variants.

Class	Spain	Costa Rica	Peru	Uruguay	Mexico
N	475	310	228	367	505
NEU	297	246	522	286	172
P	354	221	216	290	313
Σ	1126	777	966	943	990

Table 2
Distribution of tweets in the training set of the second task.

Joy	Sadness	Anger	Surprise	Disgust	Fear	Others
1270	706	600	241	113	67	2889

to compare the results (Deep Averaging Networks). 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

Two tasks have been proposed by the organizers: Task 1 - General polarity at three levels [2] and Task 2 - Emotion detection [3]. The first task consists in assigning a global polarity to tweets in three levels (**N**, **NEU** and **P**), thus collapsing the **NEU** and **NONE** classes from past editions in only one class. Several Spanish variants have been considered in this task: Spain, Mexico, Costa Rica, Uruguay and Peru. The second task is also a single-label classification task but with 7 different emotions (**joy**, **sadness**, **anger**, **surprise**, **disgust**, **fear** and **others**).

Table 1 shows the tweet distribution according to their polarity in the training set for the first task. It can be observed a bias towards the **N** and **P** classes in some Spanish variants (Spain and Mexico). In general, the **N** class is the most frequent class and the **NEU** class is the less frequent (excluding some variants like Peru or Costa Rica). In Table 2 the tweet distribution for each emotion in the training set of the task 2 is shown. In this case, there is a large bias towards the class **Others** that acts like a sink of unconsidered emotions or combinations among emotions. The less frequent class, by far, is the **Fear** class.

3. Systems

3.1. Deep Averaging Networks

We decided to use Deep Averaging Networks [4] (DAN) as baseline for this work, mainly due to their competitive performance on previous edition of TASS [5][6]. These models consist in applying feed-forward networks on top of text representations based on averaging word embeddings. Figure 1 shows an example of DAN with one hidden layer.

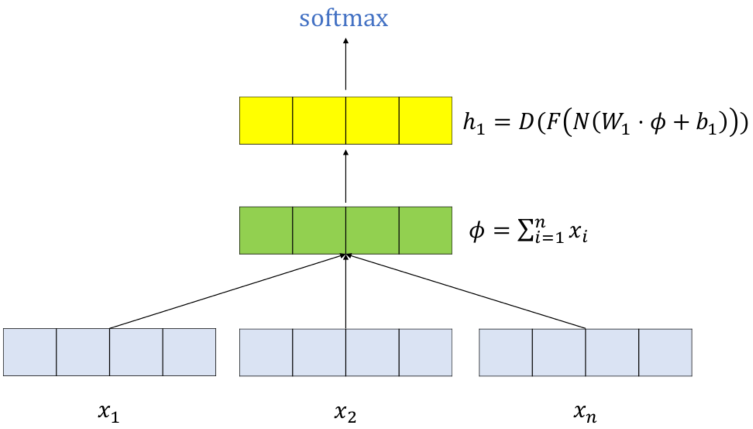

Figure 1: Deep Averaging Network: x_i is the embedding for the word i of a tweet, ϕ is the average of the word embeddings, W_1 and b_1 are the weights and bias of the hidden layer, N is the normalization strategy, F the activation function, D is the dropout and h_1 is the output of the hidden layer.

To compute the word embeddings, we used the Twitter87 model [6], that is a 300-dimensional skip-gram model [7] trained with 87 million tweets of several Spanish variants.

3.2. TWilBERT

TWilBERT is a framework for training, evaluating and finetuning BERT-based models [2] in the Twitter domain. It also includes several techniques and improvements published in recent works for the BERT architecture. Furthermore, several pre-trained models for Spanish are freely released with the framework: TWilBERT-base and TWilBERT-large. Both models were trained with 94 million of (tweet, reply) pairs in several Spanish variants.

The purpose of TWilBERT is to adapt and improve the language modeling capacity of the BERT architecture [8], based on Transformer Encoders [9], to boost the state of the art in text classification tasks on Twitter. It has several advantages in comparison to the multi-lingual version of BERT (M-BERT) for this task. First, it addresses the language dependency. M-BERT assumes that the languages used for the pre-training (104 different languages) share lexical and grammatical properties, which can induce systematic deficiencies among certain language pairs [10]. TWilBERT addresses this issue being trained from-scratch in the specific language we want to work. Second, the domain dependency. M-BERT was trained using Wikipedia texts from 104 different languages, which can degrade the performance if the target domain is very different to the domain used for pre-training. In addition, TWilBERT takes into account the coherence at tweet level by adapting the Sentence Order Prediction signal [11] to the Twitter domain. Specifically, this adaptation, called Reply Order Prediction (ROP), allows the model to learn coherence between (tweet, reply) pairs in order to improve the performance in downstream tasks that requires reasoning on pairs of tweets. Table 3 summarizes the details of the TWilBERT models in comparison to M-BERT.

[2] https://github.com/jogonba2/TWilBert

Table 3
Differences among TW-Base, TW-Large and M-BERT. L are the number of layers, A the number of attention heads in each layer, E the dimensionality of the embeddings, H the output dimensionality of each layer; and d_q, d_k and d_v are the dimensionality of the projections of the Query, Key and Value of each layer.

	M-BERT	TW-Base	TW-Large
Language	104 languages	Spanish	Spanish
Domain	Wikipedia	Twitter	Twitter
Objectives	MLM+NSP	MLM+ROP	MLM+ROP
Tokenization	WordPiece	SentencePiece	SentencePiece
Vocabulary	110k	30k	30k
Masking	Static subword	Dynamic spans	Dynamic spans
Bucketing	✗	✓	✓
L	12	6	12
A	12	6	12
E	768	768	768
H	768	768	768
d_q	64	64	64
d_k	64	64	64
d_v	64	64	64

4. Experimental Work

To carry out the experimentation for both tasks, we used DAN as baseline in order to compare the results with the TWilBERT model. TWilBERT is a pre-trained deep model while DAN is trained from scratch on the tasks. To make a fair comparison between them, a grid-search has been performed on the hyper-parameters of DAN: number of layers ({1, 2}), number of units for each layer ({64, 128}) and batch size ({8, 16, 32, 64}). Furthermore, dropout [12] was used on the output of each layer (including the input layer) with $p = 0.1$ and all the outputs were normalized by using batch normalization [13]. For TWilBERT, we did not perform any exploration of the hyper-parameters and we used directly those that obtained better results in the experimentation with the corpora of the TASS 2019. Specifically, we used TWilBERT-large with a maximum length of 128 subwords per tweet, 1e-5 as learning rate, and batches of 32 samples without gradient accumulation. All the layers of the TWilBERT model were finetuned and the vector representation of each tweet was computed as the average of the contextualized representations of the subwords inside the tweet. Both TWilBERT and DAN optimized the cross-entropy and use Adam [14] as update rule. However, TWilBERT uses weighted cross-entropy to address the imbalance among the classes of the tasks. It is important to highlight that we only used the corpora available in this edition of TASS for training both models.

Table 4 shows the results for TWilBERT and DAN, both trained with the training set, on the development set of each variant of the first task. For DAN, we only show the results obtained by the best combination of the hyper-parameters, following the aforementioned grid-search for each variant. It can be observed how the TWilBERT model outperforms the DAN model in all the variants and metrics by a large margin (between 8 and 17 points of MF_1). In average,

Table 4
Results of DAN and TWilBERT models on the development set of the task 1.

Variant	System	Acc	MP	MR	MF_1
Spain	DAN	61.10	59.17	57.51	57.99
	TWilBERT	**67.81**	**66.18**	**64.49**	**65.10**
Mexico	DAN	62.75	57.82	55.97	56.35
	TWilBERT	**70.78**	**66.08**	**66.38**	**66.20**
Costa Rica	DAN	62.56	63.25	62.24	62.49
	TWilBERT	**67.95**	**68.45**	**67.53**	**67.66**
Uruguay	DAN	60.56	59.77	59.53	59.44
	TWilBERT	**67.70**	**68.28**	**67.19**	**76.52**
Peru	DAN	58.43	55.03	57.47	55.67
	TWilBERT	**68.88**	**65.85**	**64.95**	**65.37**

TWilBERT outperforms DAN by +8.24 MF_1.

Our team presented three different runs to the competition: run-1 (DAN), run-2 (TWilBERT trained with the training set) and run-3 (TWilBERT trained with the training and development set until the epoch where the MF_1 was maximized on the development set with the run-2).

The results of our runs for each Spanish variant are shown in Table 5. These results support the competitive performance of TWilBERT, that obtains +6.42 MF_1, in average, more than DAN when it is trained with the training set (run-2) and +8.01 MF_1 when it is trained with the training and development sets (run-3). Our system TWilBERT-large, trained with all the available data of this edition, obtains the best results of the competition in the Spain, Mexico, Costa Rica and Peru variants. At this point, it is important to highlight that, to obtain these results with TWilBERT, we did not perform any exploration of its hyper-parameters. Therefore, these results could be improved by performing a more extensive experimentation.

Regarding the second task for Emotion Detection, we used the same systems than for the first task. Table 6 shows the results of DAN and TWilBERT on the development set. In this case, the results are more similar than in the previous task on the development set, with a difference of 0.5 MF_1. Also, it can be seen how the MP and MR are unbalanced for the DAN system, while for the TWilBERT system are similar between them, mainly due to the weighting of the cross-entropy.

We submitted two runs for the task 2: run-1 (DAN) and run-2 (TWilBERT trained with the training set). The results of each run in the test set are shown in Table 7. It can be observed how TWilBERT generalizes better than DAN on this set, obtaining +1.5 MF_1 in comparison to DAN, mainly due to an increment of +3.3 MR. Both systems obtained the best results of the competition.

5. Conclusions

We have proposed the use of TWilBERT for the Sentiment Analysis and Emotion Detection tasks of TASS 2020. The results obtained by our system are very promising, being the first or second ranked system in almost all the Spanish variants of the Sentiment Analysis task and the

Table 5
Results of our runs in the test set of the task 1 ("es" is the acronym for the Spain variant, "mx" for Mexico, "cr" for Costa Rica, "uy" for Uruguay and "pe" for Peru).

Run	MP	MR	MF_1
run1-cr	55.62	55.88	55.75
run1-es	58.32	58.36	58.34
run1-mx	55.60	55.75	55.67
run1-pe	59.62	53.69	56.50
run1-uy	57.81	57.68	57.74
run2-cr	63.05	62.22	62.63
run2-es	65.64	65.18	65.41
run2-mx	61.41	62.25	61.83
run2-pe	67.24	60.26	**63.56**
run2-uy	63.53	61.85	62.68
run3-cr	64.65	64.62	**64.64**
run3-es	67.27	66.96	**67.11**
run3-mx	63.70	63.19	**63.45**
run3-pe	63.51	63.27	63.39
run3-uy	66.75	64.25	**65.47**

Table 6
Results of DAN and TWilBERT on the development set of the task 2.

System	Acc	MP	MR	MF_1
DAN	67.09	63.85	51.84	54.68
TWilBERT	**67.56**	55.37	55.55	**54.84**

Table 7
Results of our runs on the test set of the task 2.

Run	MP	MR	MF_1
run-1	44.63	41.68	43.11
run-2	44.34	44.98	**44.66**

first ranked system in the Emotion Detection task. This is especially significant, considering that these results have been obtained without an exploration of the hyperparameters of the model and only a reasonable configuration was used for all the tasks.

Acknowledgments

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] J. González, L. Hurtado, F. Pla, Elirf-upv at TASS 2019: Transformer encoders for twitter sentiment analysis in spanish, in: M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, J. Carrillo-de-Albornoz, S. Montalvo, L. Chiruzzo, S. Collovini, Y. Gutiérrez, S. M. J. Zafra, M. Krallinger, M. Montes-y-Gómez, R. Ortega-Bueno, A. Rosá (Eds.), Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, volume 2421 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 571–578. URL: http://ceur-ws.org/Vol-2421/TASS_paper_2.pdf.

[2] M. García-Vega, M. C. Díaz-Galiano, M. A. García-Cumbreras, A. Montejo Ráez, S. M. Jiménez Zafra, E. Martínez-Cámara, C. A. Murillo, E. Casasola Murillo, L. Chiruzzo, D. Moctezuma, Sobrevilla, Overview of tass 2020: Introduction emotion detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS, Málaga, Spain, 2020.

[3] F. M. Plaza del Arco, C. Strapparava, L. A. Urena Lopez, M. Martin, EmoEvent: A multilingual emotion corpus based on different events, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1492–1498. URL: https://www.aclweb.org/anthology/2020.lrec-1.186.

[4] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, 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), Association for Computational Linguistics, Beijing, China, 2015, pp. 1681–1691. URL: https://www.aclweb.org/anthology/P15-1162. doi:10.3115/v1/P15-1162.

[5] J. González, L. Hurtado, F. Pla, 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., 2017, pp. 29–34. URL: http://ceur-ws.org/Vol-1896/p2_elirf_tass2017.pdf.

[6] J. González, L. Hurtado, F. Pla, 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., 2018, pp. 37–44. URL: http://ceur-ws.org/Vol-2172/p2_elirf_tass2018.pdf.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., USA, 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.

org/abs/1810.04805. arXiv:1810.04805.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[10] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, CoRR abs/1906.01502 (2019). URL: http://arxiv.org/abs/1906.01502. arXiv:1906.01502.

[11] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[13] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 448–456.

[14] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.