URJC-Team at EmoEvalEs 2021: BERT for Emotion Classification in Spanish Tweets

Jorge Alberto Flores Sánchez[1], Soto Montalvo Herranz[1], and Raquel Martínez Unanue[2]

[1] Universidad Rey Juan Carlos, Spain, jorgeflores8185@gmail.com, soto.montalvo@urjc.es
[2] Universidad Nacional de Educación a Distancia, Spain, raquel@lsi.uned.es

Abstract. This paper describes the participation of the URJC-Team in the EmoEvalEs 2021 task of the IberLEF evaluation campaign. The task consists of classifying the emotion expressed in a tweet among seven different classes of emotion. Our proposal is based on transfer learning using BERT language modeling. We train three fine-tuned BERT models finally selecting for the submitted runs two of them, along with a system that combines all the models by means of an ensemble method. We obtained competitive results in the challenge, ranking fifth. Additional work needs to be done to improve the results.

Keywords: Emotion Classification, Tweets, Deep Learning, Transformer

1 Introduction

Today it is very common that people express their emotions, feelings and opinions in social media. These emotions can be an important source of information to study people's reactions to different products, services, events or situations. For this reason, the analysis of emotions expressed in social media content has attracted researchers in the field of Natural Language Processing (NLP).

This paper describes the system presented by the URJC-Team in the EmoEvalEs challenge of IberLEF 2021 [4]. The challenge proposes to classify the emotion expressed in a tweet as one of the following emotion classes: Anger, Disgust, Fear, Joy, Sadness, Surprise or Others. Detailed information about EmoEvalEs, including a detailed description of the corpus and evaluation metrics, is provided in the shared task overview paper [5].

Our proposal is based on transfer learning using BERT language modeling. BERT fine-tuning has shown outstanding performance in a wide range of NLP tasks.

The remainder of the paper is organized as follows. Section 2 describes the data set used and the proposed system. In Section 3, the official results achieved in the challenge are presented. Finally, Section 4 summarizes the conclusions.

IberLEF 2021, September 2021, Málaga, Spain.

2 Material and Methods

2.1 Data

The EmoEvalEs dataset [6] is based on events that took place in April 2019 and related to different domains: entertainment, catastrophe, political, global commemoration, and global strike. For the task, the data were divided into dev, training and test partitions. The distribution of emotions for each dataset is shown in Table 1.

Table 1. Number of tweets by emotion.

Emotion	Dev	Train	Test
others	414	2798	814
joy	181	1227	354
sadness	104	692	199
anger	85	588	168
surprise	35	238	67
disgust	16	111	33
fear	9	65	21

To develop our proposal we have used the development and training partitions, as the test partition was later provided by the organizers to evaluate the participating systems and determine the winner of the challenge. We have merged the training and development data by blending them together to create a larger training set, which will be referred to hereafter as training data.

We randomly selected 90% of each emotion class to train the model and the remaining 10% to test it. Table 2 shows the final distribution of these data.

Table 2. Number of tweets by emotion in our train and test data.

Emotion	Train	Test
others	2890	322
joy	1267	141
sadness	715	80
anger	606	67
surprise	246	27
disgust	114	13
fear	67	7

2.2 Methods

We explore the use of Bidirectional Encoder Representations from Transformers (BERT) [2], a deep learning approach that has proven to be very successful when applied to several NLP tasks. In particular, we have experimented with a pre-trained BERT model, BETO [1], as the core for the semantic representation of the input tokens. BETO is a BERT model trained on over 300M lines of a Spanish corpus, and it is similar in size to a BERT-Base model [1].

BETO has 12 self-attention layers with 16 attention-heads each, using 1024 as hidden size. In total the model has 110M parameters. Two versions of BETO are trained: one with cased data and one with uncased data [1].

The proposed system has been implemented in Python 3.7 with Hugging-Face's transformers library [7]. Three models have been trained with different data and configuration parameters. First of all, a basic pre-processing were carried out, eliminating the special character '#'. Then it is tokenized by taking the words to subwords found in the 32k token vocabulary. Adam optimizer [3] was used with the standard parameters ($\beta 1 = 0.9$, $\beta 2 = 0.999$). We applied a linear decay function to decrease the initial learning rate to 0. And finally the max sequence length is 128 tokens.

We fixed some hyper-parameters for the different models:

- Model 1. The case model was trained with all data, batch size=32, learning rate=2e-5, epochs=4, and weight decay=0.1.
- Model 2. The uncase model was trained with all data, batch size=32, learning rate=5e-5, epochs=3, and weight decay=0.1.
- Model 3. The case model was trained with all data, but removing 30% of the "others" class since it was the majority class. Batch size=32, learning rate=2e-5, epochs=4, and weight decay=0.1.

We have submitted three different runs: one that assembles the results of the three previous models by means of a voting system, the final result being the class with the most votes, in the event of a tie, the final result will be the prediction of Model 1. The others two submitted runs with the results of models 2 and 3, respectively.

3 Results

The evaluation measures used by organizers are the following: accuracy and the weighted-averaged versions of Precision, Recall, and F1. The participant systems are ranked by the weighted-averaged F1 and accuracy measures in a multi-class evaluation scenario.

Table 3 shows the results obtained by the three runs in the challenge. The best results are for the ensemble method, this is because the predictions of the multiple models are combined taking advantage of the performance of each of them. Table 4 contains the results of the three submitted runs for each emotion. The system achieves the best results for *other*, *sadness*, *joy*, *fear* and *anger* classes. However, for the *disgust* and *surprise* classes it works badly, this is because the system confuses these emotions with others that are similar, such as *disgust* with *anger*, and *surprise* with *joy* or *others*. In addition, the small sample that is available for these classes can be affecting. Thus, the system does not have enough data to train the model and be able to differentiate between these classes.

Table 3. Results achieved by the system across the three submitted runs.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Model 2	0.688406	0.679445	0.688406	0.683092
Model 3	0.676329	0.670922	0.676329	0.672319
Voting	**0.702899**	**0.692397**	**0.702899**	**0.696675**

Table 4. Results achieved by the system across the three submitted runs for each emotion.

	Model 2			Model 3			Voting		
Emotion	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
anger	0.59	0.57	0.58	0.56	0.62	0.59	0.62	0.60	**0.61**
disgust	0.13	0.09	**0.11**	0.00	0.00	0.00	0.18	0.06	0.09
fear	0.68	0.71	**0.70**	0.64	0.76	**0.70**	0.59	0.62	0.60
joy	0.65	0.62	0.64	0.62	0.70	**0.66**	0.65	0.66	**0.66**
others	0.74	0.79	**0.77**	0.77	0.72	0.74	0.76	0.79	**0.77**
sadness	0.73	0.72	0.73	0.68	0.73	0.71	0.76	0.76	**0.76**
surprise	0.39	0.28	0.33	0.35	0.33	0.34	0.40	0.31	**0.35**

Making a comparison between Model 2 and Model 3, it can be seen that when training the model with the least amount of data for the *others* class, the performance of the model increases for the *joy* class and decreases for the *others* class. This is because with the new data distribution, the model is able to better differentiate the *joy* class over the *others* class, thus classifying a greater number of tweets correctly. Otherwise, as this data decrease is not performed for the *others* class, the performance of the *joy* class decreases, since several tweets belonging to this class are predicted by the model as *others*.

Moreover, it can be seen that although the best general results have been obtained with the voting system, there are classes like *fear* and *disgust* where this is not the case, since for Model 2, the *fear* class reaches an F1 of 0.7 and voting 0.6, and the same for *disgust* class, where Model 2 reaches 0.11 and voting 0.09.

Finally, it is important to note that although the results are slightly worse in some classes, overall robustness is gained.

On the other hand, a comprehensive comparison and ranking of the results from all the shared task participants can be found in [5]. Table 5 summarizes these results. Our system has reached position number four among the fifteen participants. Making a comparison with the best system on the accuracy metric, the difference is 2.475%, this is equivalent to the fact that the system correctly classified 41 more tweets than our system (the validation set is composed by 1656 tweets).

Table 5. Results and ranking of all competition participants.

User	**Accuracy**	**Macro averaged precision**	**Macro averaged recall**	**Macro averaged F1 score**
daveni	0.727657 (1)	0.709411 (1)	0.727657 (1)	0.717028 (1)
fyinh	0.722222 (2)	0.704695 (2)	0.722222 (2)	0.711373 (2)
HongxinLuo	0.712560 (3)	0.704496 (3)	0.712560 (3)	0.705432 (3)
JorgeFlores	**0.702899 (4)**	**0.692397 (4)**	**0.702899 (4)**	**0.696675 (4)**
hahalk	0.692029 (5)	0.679620 (6)	0.692029 (5)	0.663740 (8)
JAGD	0.685990 (6)	0.672546 (7)	0.685990 (6)	0.668407 (7)
ffm	0.684179 (7)	0.682765 (5)	0.684179 (7)	0.682487 (5)
fazlfrs	0.682367 (8)	0.664868 (8)	0.682367 (8)	0.668757 (6)
luischir	0.678140 (9)	0.658314 (9)	0.678140 (9)	0.657367 (10)
vitiugin	0.675725 (10)	0.657681 (10)	0.675725 (10)	0.661427 (9)
job80	0.668478 (11)	0.652840 (12)	0.668478 (11)	0.646085 (11)
aridearriba	0.652778 (12)	0.600479 (14)	0.652778 (12)	0.622223 (12)
Timen	0.617754 (13)	0.597877 (15)	0.617754 (13)	0.600217 (13)
QuSwe1d0n	0.536836 (14)	0.653707 (11)	0.536836 (14)	0.557007 (14)
qu	0.449879 (15)	0.618833 (13)	0.449879 (15)	0.446947 (15)

4 Conclusions

This paper describes the system presented by the URJC-Team at the EmoEvalEs 2021 task at the IberLEF evaluation campaign. Several deep-learning models were trained and ensembled to automatically detect and classify emotions expressed by people from associated events in Spanish tweets. Although it is a complex task, our system achieves good results for certain emotions and is competitive with respect to the other systems of the other participants, obtaining a difference of 2.475% with the best system of the workshop.

As future work, it is intended to carry out further experiments with BETO and other pre-trained linguistic models in order to improve the results in the task, taking into account the classes that were more difficult to detect, *disgust* and *surprise*. In addition, it might be interesting to do some preprocessing to deal with unbalanced data.

Acknowledgments

This work was supported by MCI/AEI/FEDER, UE DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32.

References

1. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: Proceedings of the Practical ML for Developing Countries Workshop at the Eighth International Conference on Learning Representations (ICLR 2020). Addis Ababa, Ethiopia (Apr 2020)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
4. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez-Carmona, M.Á., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez-Zafra, S.M., Lima, S., Plaza-de Arco, F.M., Taulé, M. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
5. Plaza-del-Arco, F.M., Jiménez-Zafra, S.M., Montejo-Ráez, A., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. Procesamiento del Lenguaje Natural **67**(0) (2021)
6. Plaza-del-Arco, F., Strapparava, C., Ureña-Lopez, L.A., Martin-Valdivia, M.T.: EmoEvent: A Multilingual Emotion Corpus based on different Events. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1492–1498. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.186
7. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. CoRR **abs/1910.03771** (2019), http://arxiv.org/abs/1910.03771