Transformers Pipeline for Offensiveness Detection in Mexican Spanish Social Media

Victor Gómez-Espinosa[1], Victor Muñiz-Sanchez[1], and Adrián Pastor López-Monroy[2]

[1] Mathematics Research Center (CIMAT), Monterrey, 66628, Mexico
[2] Mathematics Research Center (CIMAT), Guanajuato, 36023, Mexico.
{victor.gomez, victor_m, pastor.lopez}@cimat.mx
https://www.cimat.mx/

Abstract. In this paper, we describe the methodology proposed for participating in the MeOffendEs@IberLEF 2021 competition for the Subtask 3: Non-contextual binary classification for Mexican Spanish, which consists in the classification of tweets as offensive or non-offensive. We proposed a Transformers-based pipeline, consisting on a series of preprocessing steps and the use of an extended corpus, followed by an ensemble of BERT models. The proposed strategy obtained the best results on this task by ranking first place.

Keywords: Offensiveness Detection · Mexican Spanish · Transformers

1 Introduction

In the last years, there have been many initiatives in the NLP and Machine Learning community, to guide research efforts towards solutions in the automatic detection of threats and risks to the users of social networks. Those threats include aggressiveness, hate speech, harassment, racism, misogyny, among many others. For spanish language, those efforts have been promoted by academic competitions in specific tasks, such as the events organized by TASS [6,10], PAN [5] and particularly, MEX-A3T@IberLEF [1,2,7] and MeOffendEs@IberLEF [12, 13], which includes a track for aggresiveness and offensiveness identification task respectively for tweets in Mexican spanish.

Detection of offensive comments or posts in social media is not easy, because it is not depending on the presence or absence of specific words. As an example, consider the next tweets taken from MEX-A3T 2020 training corpus:

```
"No sé si guardar dinero para salir contigo o gastarlo en
pendejadas a la verga"
"Ya no saben qué verga decir, consigan una vida y sufran o algo"
```

IberLEF 2021, September 2021, Málaga, Spain.

In the first case, the tweet is non-offensive, even when it contains vulgar and rude language. The second tweet is offensive, although the language is less vulgar than the first one. Based on that, we argue that it is necessary to take into account the context in which words are used.

There are many proposals to tackle offensive and aggressive content detection in social media for spanish language, with document representations based on n-grams (word and character level), and word embeddings, with classifiers based on standard machine learning and deep learning approaches [1,3,7]. However, in the last year, there is a visible trend in the use of contextualized representations of words, such as Bi-LSTM, Bi-GRU, and Transformers-based models, such as BETO [8] with and without fine tunning [2,4,16,17]. State of the art results for Mexican spanish on this task has been reached with a bagging-like scheme, by combining different BERT models trained on different augmented datasets [11].

Similar to [11], we propose an ensemble of BERT models, but also, we use a pre-processing step in order to obtain valuable text descriptions of specialized language used in tweets, followed by an extension of the training corpus. The empirical evaluation shows that the proposed approach obtained the best results in the challenge by ranking first place. In the following sections, our proposal is explained in detail.

This document is organized as follows: Section 2 describes the dataset and the experimental settings. Section 3 describes the proposed pipeline, and Section 3.4 the experimental results. Finally, Section 4 outlines the conclusions.

2 Dataset and model settings

OffendMex corpus consists of a training set of 5060 tweets and a validation set of 76 tweets, from the total, 80% was used for training and 20% for evaluation purposes. The dataset has a length mean of 24.11 and a maximum of 60 tokens, with an unbalanced ratio of 2.66 and the offensive class as the minority.

For this task, a pre-trained BERT model on Spanish was used [8], and for the fine-tuning step for small datasets (less than 100,000), we used the exhaustive search over the recommended hyperparameters [9] and we choose the best one on the evaluation set. As a result, we used a BERT model with a training batch size of 16 for 4 epochs, and an Adam optimizer with a learning rate of 3e-5.

3 Pipeline

In this section, the proposed pipeline is described: the corpus pre-processing step, second, the extended corpus step, and finally, the BERT ensemble step.

3.1 Step 1: Pre-processing

From other classification tasks such as irony or sentiment classification has been proved that adding the tweet jargon like hashtags, emojis, and emoticons as text

descriptions improves tweet classification tasks through deep learning models like BERT [14, 15].

Our procedure is the following:

- Hashtags are split into words (see Figure 1) using the python word ninja library (https://github.com/keredson/wordninja) with a Spanish dictionary made with the Spanish fasttext vocabulary (https://fasttext.cc/docs/en/crawl-vectors.html) .
- Emojis are replaced with their text meaning in Spanish (see Figure 1) given by the python library emoji (https://github.com/carpedm20/emoji/).
- Emoticons are replaced with a text representation in Spanish similar to the emojis meanings (see Table 1).
- Words out of vocabulary are replaced with the corresponding words (see Table 1).

#YordiEnEXA Inche maricon 😂😂😂 Ahora entiendo porque eres Ro

yordi en ex a inche maricon cara llorando de risa cara llorando de risa

Fig. 1. Above: Tweet before pre-processing. Below: After pre-processing.

After pre-processing step, the maximum corpus length increases twice (see Figure 2); this is the reason why the maximum sequence length of BERT model is 128.

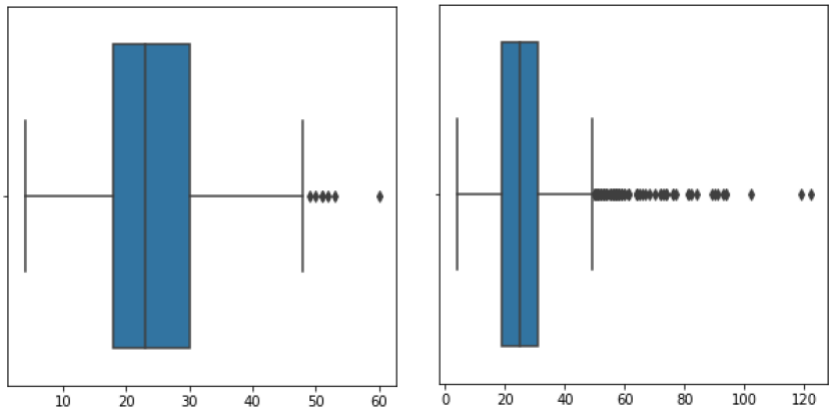

Fig. 2. OffendMex token lengths statistics. Left) before pre-processing. Right) after pre-processing.

Table 1. Expressions and their Spanish text representations for emoticons and words out of vocabulary. Note that the main idea is to describe with words the meaning of the emoticon. For example, for the emoticon :) we replace by "smiling face".

EXPRESSIONS	SPANISH
:) , =d	cara sonriendo
d: , :0 , =o	cara asombrada
:3	cara de ternura
;)	cara guiñando el ojo
u.u	cara de pena
:/	cara de confusión
-.-	cara sin expresión
¬¬	cara de desaprobación
:v	cara de repulsión
xd	cara llorando de risa
:p	cara sacando la lengua
:'v , :'c	cara llorando
:(,): , :c	cara desanimada
.l.	dedo corazón hacia arriba
hdp	hijo de puta
alv	a la verga
alm	a la madre
lol	jajajajaja
mlp	me la pelas
ptm	puta madre
pkm	poca madre

3.2 Step 2: Extend training corpus

It was demonstrated [16] that increasing the training corpus examples with other corpus labeled with a related task such as hate speech could improve model performance. In this pipeline, we choose to use hate speech and negative sentiment from the HatEval 2019 in Spanish [4] and TASS 2019 for Mexican Spanish [10] corpus, respectively.

The methodology proposed to add examples from other corpus as a way to improve model performance and reduce the unbalanced ratio is shown in Figure 3, and consists on the following three steps. In the first step the corpus must be preprocessed by the method described in section 3.1, the second step consists on training with the OffendMex 2021 corpus by the method described in section 3.3, and make inference on the HatEval and TASS corpus, and then, we select only those examples whose weights in the classifier are greater or equal to 0.95, which are added to the OffendMex corpus as offensive examples. Finally, the step three consists of training from scratch the model again. The intuitive idea of this step is to augment the training data only with those instances that could improve the classification score.

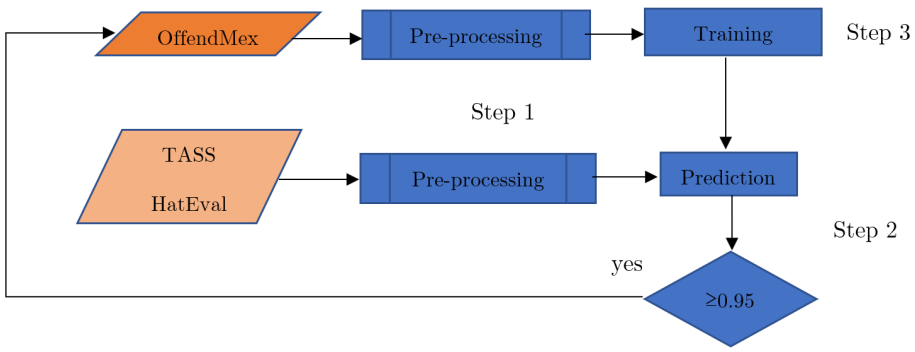

Fig. 3. Methodology to add examples from other labeled corpus related to this task.

3.3 Step 3: Bert ensemble

In order to alleviate BERT instability of fine-tuning on small samples and unbalanced datasets, it was shown [11] that using single BERTs as weak models and through an ensemble of 20 BERTs and a weighted voting scheme, which means that accumulating the softmax layer outputs and selecting the class with the maximum weight makes a more robust model (see Figure 4).

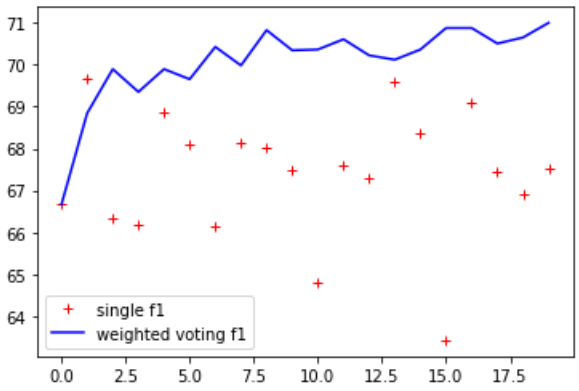

Fig. 4. F1 score for single BERT models and the ensemble (1 to 20 BERTs) with weighted voting scheme. The blue line indicates the F1 score as the BERT ensemble increases, whereas red crosses shows the score of the individual BERT added to the ensemble at each step.

3.4 Results

After following the proposed pipeline described in Section 3, we obtained the results shown in Table 2, where it can be seen that each step on the proposed pipeline helps to improve the model performance, as we expected. The best result achieved a F1 score on the evaluation set of 71.07 with a 20 BERT ensemble, pre-processing, and finally adding more examples to the training corpus.

Table 2. F1 Score on a preliminary evaluation subset (from the original training) consisting of 80% for training and 20% for validation.

20 BERT ensemble	F1(%)
Single BERT No Pre-Processing	67.19
No Pre-processing	70.28
With Pre-processing	71.00
After extending the training corpus	71.07

4 Conclusions

This work presented a pipeline of three steps for offensiveness detection on Mexican Spanish social media that effectively achieved first place on the MeOffendEs@IberLEF 2021 subtask 3 competition with a F1 score of 0.7026 on the test set. Our experimental results on the evaluation set shown that each step on the pipeline improves the model performance. We thought this pipeline could be implemented quickly and successfully in other related tasks such as aggressiveness detection.

Acknowledgements

Gómez-Espinosa thanks CONACYT for the scholarship for Master degree studies with number: 1002761.

References

1. Aragón, M.E., Carmona, M.Á.Á., Montes-y-Gómez, M., Escalante, H.J., Pineda, L.V., Moctezuma, D.: Overview of MEX-A3T at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In: 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. CEUR Workshop Proceedings, vol. 2421, pp. 478–494. CEUR-WS.org (2019)

2. Aragón, M.E., Jarquín-Vásquez, H.J., Montes-y-Gómez, M., Escalante, H.J., Pineda, L.V., Gómez-Adorno, H., Posadas-Durán, J.P., Bel-Enguix, G.: Overview of MEX-A3T at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 222–235. CEUR-WS.org (2020)
3. Aragón, M.E., López-Monroy, A.P.: Author profiling and aggressiveness detection in spanish tweets: MEX-A3T 2018. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. CEUR Workshop Proceedings, vol. 2150, pp. 134–139. CEUR-WS.org (2018)
4. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019)
5. Bevendorff, J., Chulvi, B., Sarracén, G.L.D.L.P., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska, M., Zangerle, E.: Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection. In: Hiemstra, D., Moens, M.F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) Advances in Information Retrieval. pp. 567–573. Springer International Publishing, Cham (2021)
6. Cámara, E.M., Almeida-Cruz, Y., Díaz-Galiano, M.C., Estévez-Velarde, S., Cumbreras, M.Á.G., Vega, M.G., Gutiérrez, Y., Montejo-Ráez, A., Montoyo, A., Muñoz, R., Piad-Morffis, A., Villena-Román, J.: Overview of TASS 2018: Opinions, health and emotions. 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. CEUR Workshop Proceedings, vol. 2172, pp. 13–27. CEUR-WS.org (2018)
7. Carmona, M.Á.Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H.J., Pineda, L.V., Reyes-Meza, V., Sulayes, A.R.: Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. CEUR Workshop Proceedings, vol. 2150, pp. 74–96. CEUR-WS.org (2018)
8. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
9. 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)
10. Díaz-Galiano, M.C., Vega, M.G., Casasola, E., Chiruzzo, L., Cumbreras, M.Á.G., Cámara, E.M., Moctezuma, D., Montejo-Ráez, A., Cabezudo, M.A.S., Tellez, E.S.,

Graff, M., Miranda-Jiménez, S.: Overview of TASS 2019: One more further for the global spanish sentiment analysis corpus. In: 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. CEUR Workshop Proceedings, vol. 2421, pp. 550–560. CEUR-WS.org (2019)
11. Guzman-Silverio, M., Balderas-Paredes, Á., López-Monroy, A.P.: Transformers and data augmentation for aggressiveness detection in mexican spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 293–302. CEUR-WS.org (2020)
12. 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)
13. Plaza-del-Arco, F.M., Casavantes, M., Escalante, H., Martin-Valdivia, M.T., Montejo-Ráez, A., Montes-y-Gómez, M., Jarquín-Vásquez, H., Villaseñor-Pineda, L.: Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. Procesamiento del Lenguaje Natural **67**(0) (2021)
14. Pota, M., Ventura, M., Catelli, R., Esposito, M.: An effective bert-based pipeline for twitter sentiment analysis: A case study in italian. Sensors **21**(1) (2021)
15. Singh, A., Blanco, E., Jin, W.: Incorporating emoji descriptions improves tweet classification. 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. 2096–2101. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
16. Tanase, M., Zaharia, G., Cercel, D., Dascalu, M.: Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 236–245. CEUR-WS.org (2020)
17. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç.: SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 1425–1447. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020)