Evaluation of Intermediate Pre-training for the Detection of Offensive Language

Segun Taofeek Aroyehun and Alexander Gelbukh

CIC, Instituto Politécnico Nacional Mexico City, Mexico
aroyehun.segun@gmail.com & gelbukh@gelbukh.com

Abstract. This paper presents an evaluation of intermediate pre-training for the task of offensive language identification. We leverage recent advances in multilingual contextual representation and fine-tuning of pre-trained language models. We compare the performance of a pre-trained language model adapted for the social media domain and another that was further trained on multilingual sentiment analysis data. We found that the intermediate pre-training steps prior to fine-tuning on the target task yield performance gains. The best submissions by our team, NLP-CIC, achieved first and second place on the non-contextual Spanish (Subtask 1) and Mexican Spanish (Subtask 3) subtasks of the MeOffendEs-IberLEF 2021 shared task respectively.

Keywords: XLM-RoBERTa · Social Media · Spanish · Mexican Spanish · Offensive Language Identification · Sentiment Analysis.

1 Introduction

The purpose of social media is for information exchange. This involves interactions among users on the various social media platforms. During these interactions, users often show unhealthy and anti-social behaviour such as insults and personal attacks. This kind of behaviour hampers meaningful conversations at the least and can cause harm to individuals, groups, and the society at large. Natural language processing research can help in identifying offensive language to help reduce incidences of unacceptable behaviour. Research into this problem has gained attention especially in English language. This is attributable to availability of labeled data and pre-trained word embeddings and language models. Recently, there has been a number of shared tasks with a focus on languages other than English. One example is the IberLEF shared task series on offensive language identification in Mexican Spanish. For the 2021 edition [8] the tasks include MeOffendEs [11], a track on offensive language identification on several social media platforms. The challenge includes datasets for Spanish and Mexican Spanish.

IberLEF 2021, September 2021, Málaga, Spain.

Multilingual language models (LM) are becoming a popular area of focus with the transformer architecture [13] which makes it possible to combine text written in different languages to learn a single multilingual representation. There are several successful instances of this approach in multilingual BERT [6], XLM-RoBERTa [5], and recently multilingual T5 [15]. However, these pre-trained multilingual LMs mainly cover domains with text written in consistent and formal style in contrast to social media text which are noisy, irregular and informal. To adapt LMs to a specific domain, the authors of [4, 7] showed that using the transformer architecture and its pre-trained weights, a domain-specific model can be derived by continuing pre-training on text specific to the domain of interest. For the social media domain, [9] and [3] are examples of this adaptation in the monolingual English setting. For the multilingual case, the authors of [2] introduced an adaptation of XLM-RoBERTa to multilingual twitter text. XLM-RoBERTa was further trained with the masked language modeling objective on twitter text (about 12GB) in over 30 languages. Furthermore, this LM was trained on a unified collection of sentiment analysis data in eight languages with the goal of demonstrating the effectiveness of the multilingual LM trained on twitter text.

In this paper, our focus is to evaluate the effectiveness of the pre-trained LMs on the identification of offensive language in tweets written in Spanish and Mexican Spanish. We examine the effect that intermediate pre-training on sentiment analysis, a la [10, 12], has on offensive language identification.

2 Methodology

Task. We address the non-contextual classification of offensive language in tweets written in Spanish (Subtask 1) and Mexican Spanish (Subtask 3). For Subtask 1, the task is to classify comments written in Spanish using only the textual content into one of four categories: Offensive where the target is a person (OFP); Offensive where the target is a group of people (OFG); non-offensive, but with inadequate language (NOM); non-offensive (NO). This subtask also assess the agreement between the confidence of model predictions and the confidence of human annotators. Subtask 3 is a binary classification of tweets written in Mexican Spanish. It requires predicting whether a comment is offensive or not.

Data. The MeOffendEs 2021 shared task [11] provides two corpora, OffendEs and OffendMEX, which are collections of messages on social media platforms in Spanish and Mexican Spanish annotated with labels indicating offensiveness. The generic Spanish data consist of labeled comments focusing on popular young Spanish influencers collected from different social media platforms (YouTube, Instagram, and Twitter). The Mexican Spanish dataset was collected from Twitter and manually labeled for offensiveness. In addition, metadata for each comment is provided for the classification in the contextual tracks of the competition. We only participate in the non-contextual tracks in both languages. Table 1 provides details of the Mexican Spanish dataset. Also, the details of the Spanish dataset is in Table 2.

Table 1. Details of the datatset for Mexcian Spanish

Class	Train	Dev.	Test
0	3679	35	–
1	1381	41	–
Total	5060	76	2183

Table 2. Details of the datatset for generic Spanish

Class	Train	Dev.	Test
NO	13212	64	–
NOM	1235	10	–
OFP	2051	22	–
OFG	212	4	–
Total	16710	100	13606

We perform minimal pre-processing of the data in our experiments as it was reported in [1] that extensive pre-processing tends to hurt performance of pre-trained LMs. Hence, we normalize the text by converting user mentions and web links to *@USER* and *URL*. We also replace multiple consecutive whitespaces with a single one and punctuation marks are surrounded by a single whitespace character on both sides. Then, the sequence of text is tokenized with the subword tokenizer provided with the XLM-RoBERTa model, which is a Sentence Piece model (using a unigram language model) with a vocabulary size of 250K [5]. We set the maximum sequence length to 128 subword tokens.

Fine-tuning. We use the huggingface transformers library [14] for the experiments. We add a linear prediction layer on top of the pooled output of the last transformer layer and optimize this layer jointly with the pre-trained layers. We optimize the model using Adam (without bias correction) with a batch size of 128 on a single Nvidia V100 GPU (32GB) and a maximum learning rate within the range of 1e-5 to 5e-5. We use a warmup ratio of 0.1 and set the maximum number of epochs to 10 with earlystopping on the validation performance metric (micro F1) using a patience of 2 evaluation runs. We evaluate the performance of the model every 20 steps on the validation set. Furthermore, we employ three regularization techniques: weight decay with a factor of 0.01, dropout applied to the pooled output of the last transformer layer with a probability of 0.2, and label smoothing with a factor of 0.1. Our submissions vary in their use of the regularization approaches. Details of the settings for each submission are in Table 3. The configurations are based on XLM-twitter [1] and XLM-twitter-sentiment [2] pre-trained models introduced in [2]. For all experiments, we set the random seed to 42. On average, the fine-tuning procedure takes about 900 seconds (wall time) for the Spanish task and approximately 600 seconds (wall time) for the Mexican Spanish task.

[1] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base
[2] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

Table 3. Model configuration of submissions made to the non-contextual Mexican Spanish (MX_ES) and generic Spanish (ES) tracks. LR is the maximum learning rate and LSF is the label smoothing factor.

	Language	Model	LR	Dropout	LSF
Submission-I	MX_ES	XLM-twitter-sentiment	1e-5	0.2	0.1
Submission-II	MX_ES	XLM-twitter-sentiment	2e-5	0.2	0.0
Submission-III	MX_ES	XLM-twitter	2e-5	0.2	0.1
Submission-IV	MX_ES	XLM-twitter	2e-5	0.2	0.0
Submission-I	ES	XLM-twitter	4e-5	0.2	0.0
Submission-II	ES	XLM-twitter-sentiment	2e-5	0.2	0.0
Submission-II	ES	XLM-twitter	3e-5	0.2	0.1

3 Results

The scores of our submissions on the development and test sets for Subtask 1 (generic Spanish) are in Table 4. Three submissions are allowed for this subtask. Submission-II which is based on the XLM-RoBERTa model trained on both multilingual twitter text and sentiment analysis dataset achieved our best submission out of the three. The model also has the least mean squared error, an indication of greater agreement with the confidence of human annotators. The overall ranking showed that this system is the best on the competition for the non-contextual classification in Spanish. Figure 1 shows the confusion matrix of the best model (Submission II) on the validation dataset for Subtask 1. It can be observed that most of the mistakes on the validation data occurs where the model predicts non-offensive (NO) when the comments are actually offensive to a person (OFP). Also, the model performs poorly on the offensive to a group category (OFG). It makes the correct prediction on 1 out of 4 examples in the validation set.

Table 5 presents the results received by the NLP-CIC team on the leaderboard on the unseen test set for Subtask 3 (Mexican Spanish). The maximum number of submissions for this task is five. It can be observed that the XLM-Roberta model that has been further trained on twitter data and a collection of sentiment analysis datasets in eight languages (Submission-I) that we fine-tune on Subtask 3 dataset has the highest score out of our four submissions. Also, the results show that label smoothing was beneficial for this task. On the overall ranking for the competition, this system is in second place. In Figure 2, the confusion matrix provides an overview into the best model performance (Submission I) across the two classes. The model predicts the non-offensive label (NO) on 8 examples when the true label is offensive (OFF) compared to the converse where the model predicts the offensive label on 1 example when the true label is non-offensive. It shows that the model is relatively better at identifying the non-offensive category on the validation dataset. This can be linked to the number of examples for the non-offensive category which is about three times more than the offensive category in the training set.

Table 4. Performance scores on the development and test sets for submissions to the generic Spanish non-contextual track. MSE is the mean squared error computed on the model prediction confidence for the test set against the confidence of the annotators. Best scores are in bold and second best scores are underlined.

		Micro-averaged Test Scores			
	Dev. Micro F1	Precision	Recall	F1	MSE
Submission-I	**0.8700**	0.8430	0.8430	0.8430	0.0330
Submission-II	0.8500	**0.8816**	**0.8816**	**0.8816**	**0.0231**
Submission-III	<u>0.8600</u>	<u>0.8493</u>	<u>0.8493</u>	<u>0.8493</u>	<u>0.0313</u>

Table 5. Performance scores on the development and test sets for the Mexican Spanish non-contextual track. The highest scores are in bold and the second highest are with underline.

		Macro-averaged Test Scores		
	Dev. Micro F1	Precision	Recall	F1
Submission-I	**0.8816**	0.7550	**0.6407**	**0.6932**
Submission-II	<u>0.8553</u>	**0.8183**	0.5756	<u>0.6758</u>
Submission-III	0.8421	0.7800	<u>0.5872</u>	0.6700
Submission-IV	**0.8816**	<u>0.7867</u>	0.5834	0.6700

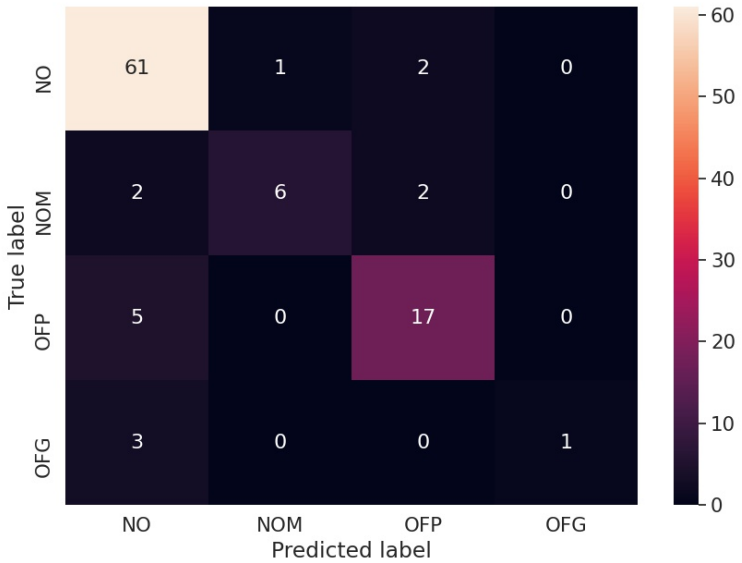

Fig. 1. Confusion matrix of the best model (Submission II for ES) on the validation dataset for the generic Spanish task. NO means non-offensive; NOM means non-offensive, but with inadequate language; OFP means offensive where the target is a person; OFG means offensive where the target is a group of people.

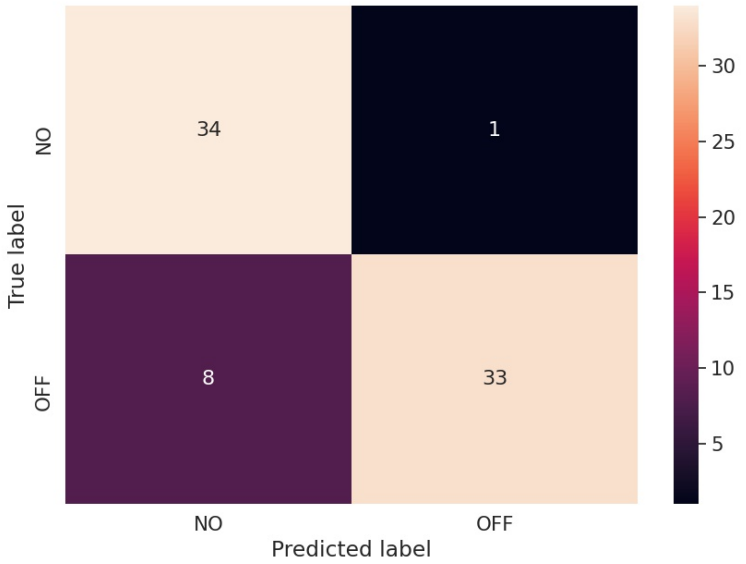

Fig. 2. Confusion matrix of the best model (Subission I for MX_ES) on the validation dataset for the Mexican Spanish task. NO refers to non-offensive; OFF refers to offensive.

We observed that overall, the scores on the Spanish dataset is far higher than the Mexican Spanish dataset even though it's a binary classification task. We think that the amount of data available for the Spanish task is a factor for this difference in performance. The consistent performance of the model that includes sentiment analysis as part of the pre-training for both tasks confirms our hypothesis that sentiment analysis can be beneficial for detecting offensiveness.

4 Conclusion

We address the task of offensive language identification in Spanish and Mexican Spanish using a pre-trained language model adapted for the twitter domain. We found that a further training on multilingual sentiment analysis is beneficial to the task. In addition, label smoothing proved useful on the Mexican Spanish dataset. The best systems submitted by our team, NLP-CIC, achieved first place on the non-contextual Spanish task and second place on the non-contextual Mexican Spanish task.

In the future, we will like to examine whether a model trained on Spanish data can be seamlessly transferred to Mexican Spanish for this task and vice versa. Our models only use textual content, it is very likely that the addition of metadata can improve their performance.

Acknowledgements

Thanks to the competition organizers for their support. The authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies.

References

1. Aroyehun, S.T., Gelbukh, A.: NLP-CIC at HASOC 2020: Multilingual Offensive Language Detection using All-in-one Model. In: FIRE (Working Notes). pp. 331–335 (2020), http://ceur-ws.org/Vol-2826/T2-31.pdf
2. Barbieri, F., Anke, L.E., Camacho-Collados, J.: XLM-T: A Multilingual Language Model Toolkit for Twitter. arXiv preprint arXiv:2104.12250 (2021)
3. Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., Neves, L.: TweetEval: Unified benchmark and comparative evaluation for tweet classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1644–1650. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.148, https://www.aclweb.org/anthology/2020.findings-emnlp.148
4. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1371, https://www.aclweb.org/anthology/D19-1371
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://www.aclweb.org/anthology/2020.acl-main.747
6. 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
7. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8342–8360. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.740, https://www.aclweb.org/anthology/2020.acl-main.740
8. 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)

9. Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.2, https://www.aclweb.org/anthology/2020.emnlp-demos.2
10. Phang, J., Calixto, I., Htut, P.M., Pruksachatkun, Y., Liu, H., Vania, C., Kann, K., Bowman, S.R.: English intermediate-task training improves zero-shot cross-lingual transfer too. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. pp. 557–575. Association for Computational Linguistics, Suzhou, China (Dec 2020), https://www.aclweb.org/anthology/2020.aacl-main.56
11. Plaza-del-Arco, F.M., Casavantes, M., Jair Escalante, H., Martín-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)
12. Pruksachatkun, Y., Phang, J., Liu, H., Htut, P.M., Zhang, X., Pang, R.Y., Vania, C., Kann, K., Bowman, S.R.: Intermediate-task transfer learning with pretrained language models: When and why does it work? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5231–5247. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.467, https://www.aclweb.org/anthology/2020.acl-main.467
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., 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), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
14. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://www.aclweb.org/anthology/2020.emnlp-demos.6
15. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)