Bi-Ensembles of Transformer for Online Bilingual Sexism Detection

Emilio Villa-Cueva[1,*,†], Fernando Sanchez-Vega[1,2,3,†] and Adrián Pastor López-Monroy[1,†]

[1]*Mathematics Research Center (CIMAT), Jalisco S/N Valenciana, 36023 Guanajuato, GTO México.*

[2]*Consejo Nacional de Ciencia y Tecnología (CONACYT), Av. de los Insurgentes Sur 1582, Benito Juárez, 03940, CDMX, México.*

[3]*El Colegio de México (COLMEX), Picacho Ajusco 20, Tlalpan, 14110, CDMX, México.*

Abstract

We present our approach for Task 1 of EXIST 2022, which consisted in identifying sexism in both English and Spanish tweets. We propose a bi-ensemble that uses a combination of two transformers ensemble, one built with models pretrained in Spanish and the other with models pretrained in English, both of them based either on the RoBERTa or the BERT architecture. The bi-ensemble method merges two previous approaches: ensembles with two architectures and ensembles with several trained models of the same architecture. Intuitively, this approach can take advantage of the knowledge of two different architectures and the robustness of ensembles to improve classification performance in this task. Our experimental results show that our approach is an effective method for the task. In the forum's evaluation, our submissions scored second team's place in task 1 with a difference of 0.0038 in F1.

Keywords

Sexism detection, Transformer Ensemble, Bi-Ensemble Transformer, Multilingual Ensembles

1. Introduction

Nowadays large communities interact within social media platforms such as Twitter, where users are exposed to a large variety of harmful content. Sexism is one of the most important because it could have a significantly negative impact on users. For example, objectification —a type of sexism— can eventually lead to significant mental health risks such as eating disorders, unipolar depression and sexual dysfunction [1]. Furthermore, the anonymity of Twitter can encourage people to display an even greater sexist behavior: a study presented in [2] asked participants to retweet or share tweets with sexist content and then conducted a set of tasks intended to identify sexist behaviors. The authors concluded that users shielded by the anonymity of a social media profile were more likely to display hostile sexism than non-anonymous users.

IberLEF 2022, September 2022, A Coruña, Spain

*Corresponding author.

†These authors contributed equally.

evillacueva@gmail.com (E. Villa-Cueva); fernando.sanchez@cimat.mx (F. Sanchez-Vega); pastor.lopez@cimat.mx (A. P. López-Monroy)

0000-0003-4998-7929 (E. Villa-Cueva); 0000-0002-8533-2818 (F. Sanchez-Vega); 0000-0003-1018-4221 (A. P. López-Monroy)

Surprisingly, little work has been devoted to the problem and the online detection remains a difficult unsolved task. Sexism detection is challenging because it can present itself in a seemingly harmless manner, such as jokes or comments of appearance. EXIST 2021 is a shared task introduced in IberLEF 2021, whose goal is to aid the identification of sexism in a *broad* sense. Posts from Twitter and Gab.com in English and Spanish were collected and labeled under the supervision of experts in gender issues [3]. These posts could be a report or description of sexist behavior, an implicitly sexist comment or an explicitly misogynistic post. This *broad* labeling makes the EXIST dataset the first of its kind. For EXIST 2022 [4], around 1058 tweets have been added to the dataset following the same annotation procedure.

The best results in EXIST 2021 were obtained using models based on the Transformer architecture [5] such as BERT [6], RoBERTa [7] or XLM-R[8]. In particular, Transformer Ensembles were used by 5 out of 33 participants, including the top-performing submission [9], whose approach mainly used ensembles to *fuse* the results of models that worked well in different languages to address the bilingual nature of the dataset. Other of the advantages of using ensembles is that they aid in reducing the known instability of BERT [10] when it is fine-tuned in a relatively small dataset.

In this work, we present CIMAT team submission for EXIST 2022. In contrast from previous approaches, our proposal consists in a *bi-ensemble model* that uses a *pair* of ensemble models, each one composed by a set of 10 instances of a given model pretrained in a specific language domain. This means a total of 20 models inside the bi-ensemble model, where a weighted voting strategy is then used to perform the final decision. The use of 10 models per domain is meant to reduce the variance in performance, while the use of two different ensembles helps in classifying correctly the tweets both in English and Spanish. The idea is that a set of models might be better for Spanish tweets and the other for English tweets. One of the used models was RoBERTuito [11], which we found to be reliable for this task because of its outstanding performance in bilingual datasets.

Experimental results show that our approach improves over single-models and traditional single-model ensembles. Moreover, our submitted proposals for task 1 of EXIST 2022 ranked 3rd, 4th and 5th out of 47. This work is structured as follows: Section 2 briefly introduces the models used in our proposal and expands on the use of ensembles for stabilizing and improving BERT performance and to tackle bilingual problems, Section 3 describes our proposed bi-ensemble method, Section 4 further describes the EXIST dataset and briefly explains the technical details of our implementation, Section 5 contains the results of our experiments in an in-house evaluation partition and the final results in the test set of the competition, finally, Section 7 concludes this article.

2. Related Work

2.1. BERT for social media tasks.

The success of domain-specific pretrained models has paved the way for BERT-based models that are pretrained in Twitter corpora. BERTweet [12] uses the BERT architecture and has been pretrained on 850M English tweets, using the RoBERTa pretraining procedure. It is designed to be used in downstream tasks that contain Twitter text, where it indeed surpasses BERT for

these kind of tasks.

RoBERTuito [11] aims to do the same for Spanish tweets: its authors pretrained a RoBERTa model with a corpora of around 500M tweets, mainly in Spanish (they allowed other languages to be in the pretraining data, estimating that around 92% is in Spanish and the rest of the corpora contains other languages) using the same pretraining configuration of BERTweet. Their results outperform other state-of-the-art models in Spanish social media tasks, where performance evaluations in Spanish-English datasets achieve outstanding results as well. As for English datasets, it outperforms multilingual models such as mBERT [6] and XLM-R [8] for classification tasks, but is not as effective as BERTweet, a model designed to work with tweets in English.

Although RoBERTuito's performance in English tasks suggests that it may work as a bilingual model, the unequal proportion of Spanish to English tweets used in the pretraining stage make it a model that is mostly tailored for Spanish tasks. mBERT (or multilingual-BERT) is a truly multilingual model that has been pretrained on more than 104 languages, nevertheless, this pretraining has been carried out using corpora from Wikipedia, whose language domain is significantly different from that of Twitter. This can be relieved by carrying out a simple Domain Adaptation following the AdaptaBERT procedure presented in [13], transitioning the language domain of mBERT to a specific language domain.

We conducted this adaptation on mBERT to shift its language domain to a social media domain similar to the EXIST dataset. We constructed our target domain data using tweets obtained from the *Sentiment140* dataset [14] in the case of English tweets and from the *TwitterSentimentDataset* dataset [15] for Spanish tweets, resulting in a multilingual BERT adapted to a Twitter language domain. From here onward, we will refer to mBERT with our adaptation as mBERT-T.

2.2. BERT stability issues and ensembles

Several works [10, 16, 17] have pointed out that fine-tuning BERT for a downstream task can lead to a high variance in performance, specially when the training data is relatively small. The general agreement is that random seeds affect performance and different seeds can result in very different performance results using the same dataset and the same model. This is true for the seeds used in initializing the weights of the classification layers and for those used in the data ordering at the fine-tuning stage [10].

In this paper, our proposal follows some of the ensemble ideas presented in [18, 19], which consisted in an single-model ensemble with weighted voting. Their idea is simple but effective, a set of N BERT models are individually fine-tuned using the training data. The classification is performed over the *test* data by aggregating the outputs of the softmax layer of the N BERT models, then the class with the highest value is selected as the prediction (a graphical description of this system is illustrated in Figure 1). Intuitively, this ensures that the selected class is the one that *most of the models* are *most confident* on. As [18] found, the performance of this approach improves as more instances of the model are incorporated to the ensemble, but reaches a plateau at around $N = 10$ models, this is why we use this value for our implementation in Section 3.

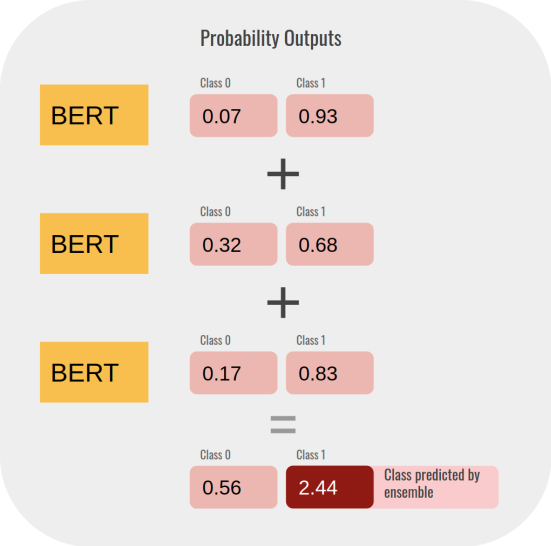

Figure 1: Ordinary weighted voting system, usually employed in competitions such as [18, 19]

2.3. BERT models for bilingual classification through ensembles

For the task of bilingual sexism identification the use of ensembles is not new, in EXIST 2021 several teams used ensembles as part of their proposal. For example, the first place was obtained by [9] through the use of an ensemble of different BERT-based models pretrained in English, Spanish and multilingual. This idea was implemented by other participants as well, who approached the problem using similar configurations with some variations, hypothesizing that an ensemble of models pretrained in different languages could *fuse* its knowledge to make better predictions in a bilingual dataset such as EXIST.

Notwithstanding the fact that the latter two ensemble strategies were previously used, they have never been combined before for this task, that is precisely the contribution of this paper.

3. A *bi-ensemble* technique for robust bilingual classification

Previously, researchers have used ensembles of more than one different Transformer [9], or ensembles of several Transformers of the same type [18, 19]. However, in this paper we devised a novel bi-ensemble that combines the best of these strategies. This new bi-ensemble is built by combining two different pretraining domains in an ensemble configuration with weighted voting. In a few words, instead of using an ensemble of N instances of the same model, or an ensemble composed by two different models (as explained in sections 2.2 and 2.3 respectively), we combine these approaches by using 20 models composed by two sets of 10 instances of each one. As shown in Figure 2, this aims to create a pipeline that is robust to both the instability of BERT and the bilingual nature of the data. This is thanks to the two models pretrained in different languages as the *building blocks* of each set. Finally, models are trained individually and their predictions are aggregated through weighted voting at the moment of inference.

We observed that for tasks related to Spanish tweets RoBERTuito significantly outperforms other pretrained models, thus, we evaluate it in different pair configurations along with other model aimed at predictions in English, resulting in the following three proposed configuration pairs:

1. RoBERTuito + BERT-en
2. RoBERTuito + BERTweet
3. RoBERTuito + mBERT-T

We fine-tune 15 individual runs for each model (RoBERTuito, BERT-en, BERTweet y mBERT-T). Then, we evaluate their F1-macro performance on the *training set*, and keep 10 out of 15 that yield the highest scores to use them in the ensembles. This selection effectively keeps the models that perform better in that specific partition, we expect that in an ensemble configuration this will help the overall performance on unseen data.

Predictions over the test set are computed using weighted voting by aggregating the probability outputs of the classification layer for the 20 models. The label with the largest probability sum will be the prediction of the ensemble for that tweet (See Figure 1).

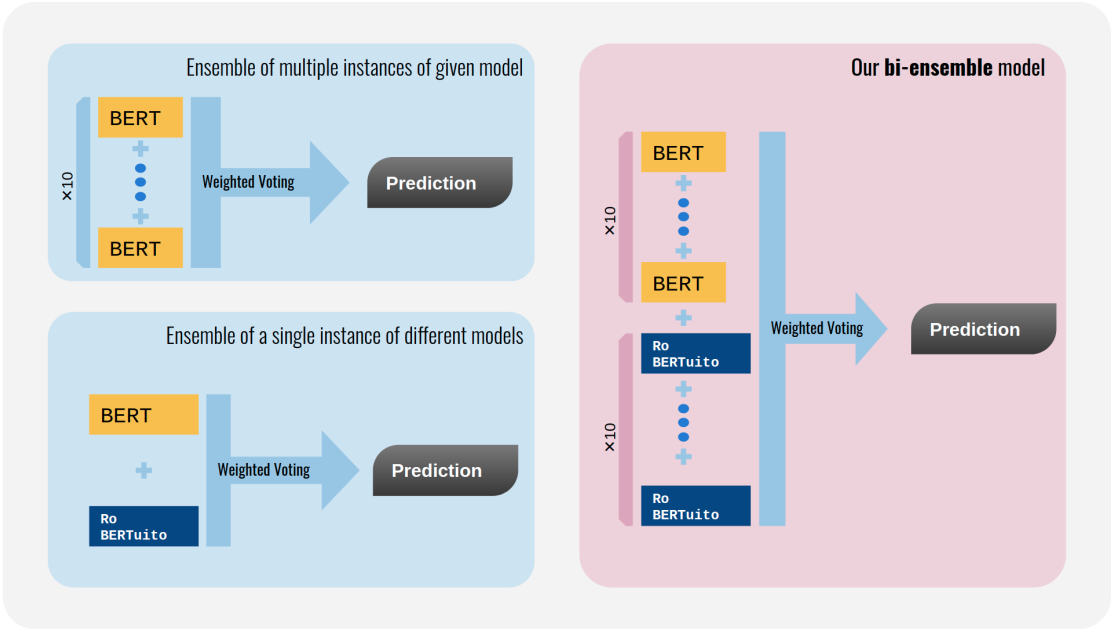

Figure 2: Our bi-ensemble configuration, the top-left panel shows an ensemble of 10 BERT models that improves prediction stability, the panel on the bottom-left displays a BERT(English)+RoBERTuito(Spanish) ensemble for English and Spanish bilingual tasks and in the right panel we illustrate our approach merging these two ideas in a bi-ensemble using a total of 20 models.

4. Dataset and Experimental Settings

The EXIST 2021 dataset [3] contains a collection of tweets labeled to classify sexism from a broad perspective. Authors extracted tweets from a variety of accounts filtering by a set of hashtags and expressions, these tweets were then labeled by 5 annotators using a majority vote and disagreements in the votes were reviewed by two experts. The labels are set according to two tasks:

- Task 1: Binary label that indicates if sexism is present in the tweet. (*sexist* or *non-sexist*)
- Task 2: Multiclass label that defines the type of sexism present (*Ideological and inequality, Stereotyping and dominance, Objectification, Sexual violence, Misogyny and non-sexual violence*)

The EXIST 2021 dataset resulted in 6977 and 3386 tweets for training and test respectively, adding to more than 10 thousand tweets. For EXIST 2022, the entire EXIST 2021 dataset (*train* and *test* partitions) was set as the training data, and for the test data 1058 new tweets were extracted and labeled using the same criteria and a very similar annotation procedure. Our submission for EXIST2022 focuses on task 1, the binary classification problem.

We preprocess the tweets using a fairly standard procedure, which is shown in Figure 3. Finally, for our experiments we fine-tuned each of the models for the classification task for 3 epochs, using a batch size of 16 and a learning rate of 2×10^{-5} for RoBERTuito and 1×10^{-5} for the rest of the models. All of our models were trained on an NVIDIA RTX Titan GPU using Pytorch and the transformers library.

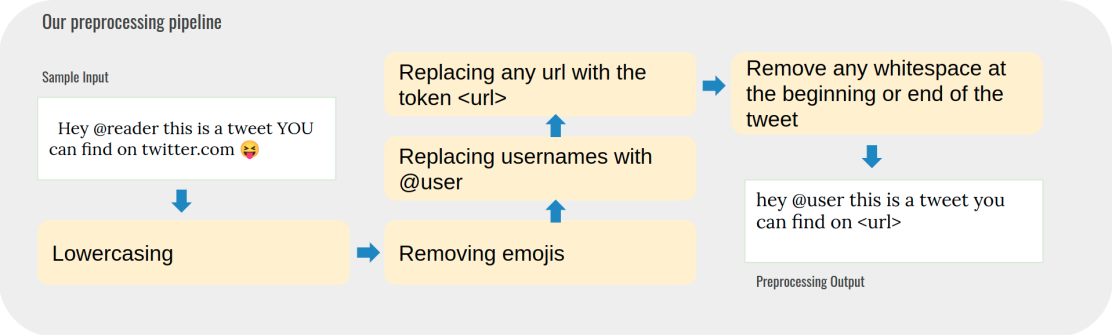

Figure 3: Our tweet preprocessing pipeline.

5. Evaluation Results

In this section we evaluate the performance of our proposals. Section 5.1, presents our experiments on an in-house evaluation partition, whereas in Section 5.2 our results in the EXIST 2022 test set.

5.1. Preliminary evaluation with in-house partitions

The purpose of doing a preliminary evaluation is to observe the performance of different models and ensembles using an *in-house* data partition. The aim was to decide the best configurations for our bi-ensemble model for the test set of EXIST 2022.

For the in-house *train* and *test* partitions, we randomly split the complete (*train* + *test*) EXIST 2021 data in two parts: 85% for training and 15% for validation. We later use it for fine-tuning and evaluation of our models respectively, using the F1-macro as our performance metric.

In our first experiment, we evaluated the performance of different individual models to observe which ones would be adequate for the bi-ensemble model. Their F1-macro scores are shown in Table 1, where RoBERTuito clearly outperforms other models due to its almost bilingual capabilities mentioned in Section 2.1. Based on these results, we selected RoBERTuito as our main model for our three bi-ensemble configurations. Interestingly, BERTweet fails to converge in some runs, affecting the average performance and resulting in a very large variance.

Table 1
Average performance of individual models over 10 runs, we can observe that RoBERTuito outperforms every other model by a large margin.

Model	Macro F1 (mean)	Macro F1 (std dev)
BERTweet	71.79	9.62
mBERT	76.98	0.63
BERT-en	75.29	0.48
mBERT-T	77.81	0.38
RoBERTuito	**79.95**	0.46

Then, we conduct a second experiment to evaluate the performance of single-ensemble and bi-ensemble configurations, for this we fine-tuned 15 instances of each model and selected 10 using the selection procedure explained in Section 3. These experiments were carried out to evaluate if the bi-ensemble approach improved performance and also to determine if the selection procedure was an useful resource. The results of these experiments are shown in the Table 2. Comparing single-model ensemble results, we can see that for most cases their overall performance is improved compared to their average individual performance. Also, the convergence problems of BERTweet are reduced. We also noted that the model selection is successful further improving the F1 macro score using smaller ensembles. Interestingly, the best performing combination is *RoBERTuito+BERT(en)*, and not *RoBERTuito+BERTweet*, which we initially though would score the highest due to the domain-specific pretraining of BERTweet compared to BERT, we hypothesize this could be related to the instability of BERTweet in this dataset.

The results of these experiments confirm that our proposed approach is effective in the EXIST 2021 dataset and we expected a similar performance for the EXIST 2022 test set.

Table 2
Performance of different ensemble configurations in in-house evaluation. We can observe that the all bi-ensemble ensembles achieve a relevant improvement over the single-ensemble models

Models in ensemble	Macro F1 ($\times 15$ without selection)	Macro F1 ($\times 10$ with selection)
BERT-en	76.338	75.991
BERTweet	77.030	77.083
mBERT-T	77.812	77.706
RoBERTuito	80.411	80.450
RoBERTuito+ mBERT-T	80.148	80.548
RoBERTuito+BERTweet	81.041	81.042
RoBERTuito + BERT-en	81.107	**81.166**

5.2. Task Evaluation

In this section we briefly present the performance metrics of our bi-ensemble configurations in the EXIST 2022 evaluation. Our submissions ranked 3th, 4th and 5th place, obtaining the scores shown in the Table 3. We note that the order our models are ranked coincides with their order in our in-house evaluation partition.

Table 3
Performance of different ensemble configurations in EXIST 2022 **test set**.

Submitted (Ranking)	Accuracy	Macro F1
RoBERTuito+BERT-en (3rd)	**79.49**	**79.40**
RoBERTuito+BERTweet (4th)	79.11	79.04
RoBERTuito+mBERT-T (5th)	78.83	78.77

6. Ethical Issues

It is important to point out that the performance improvement of our proposal comes at a cost. Training and deploying a bi-ensemble is more costly compared to using single models because we are effectively using 20 times more computational resources at the moment of training and inference, resulting in a less environmentally friendly solution.

The proposed solution for the task of bilingual sexism detection presented in this work extracts the patterns from the training data, so the classification obeys the labeling criteria in this particular corpus. Applying this type of learning-based solution to a cultural context different from the one assumed by the annotators requires a careful review by experts in language, culture, and human rights.

7. Conclusion

In this work we have outlined the main aspects of our approach for the IberLEF shared task EXIST 2022. To the best of our knowledge, this is the first time that a bi-ensemble of BERT-based models is proposed for the task of sexism detection. Evaluation results show that our proposal achieves a significant improvement over a traditional fine-tuning of state-of-the-art individual models. This was possible thanks to the advantage of ensembles in successfully tackling two critical issues: classification in a bilingual setting and fine-tuning instability, which has been pointed out by other authors.

By using this —easy to implement, but effective— procedure, our submissions obtained 3rd, 4th and 5th place out of 47 in the test set, suggesting that our method is competitive and might be an effective solution for the task of bilingual broad-sexism detection compared to other submitted proposals.

Acknowledgments

The authors thank *Consejo Nacional de Ciencia y Tecnología* (CONACYT), *Centro de Investigación en Matemáticas* (CIMAT) and *Instituto Nacional de Astrofísica, Óptica y Electrónica* (INAOE) for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies (*Laboratorio de Supercómputo: Plataforma de Aprendizaje Profundo*) with the project *"Identification of Aggressive and Offensive text through specialized BERT's ensembles"* and CIMAT Bajio Supercomputing Laboratory (#300832). Sanchez-Vega would like to thank CONACYT for its support through the Program *"Investigadoras e Investigadores por México"* by the project *"Desarrollo de Inteligencia Artificial aplicada a la prevención de violencia y salud mental."* (ID. 11989, No. 1311) and the COLMEX Interdisciplinary Data Science Program (Open Society Grant).

References

[1] B. L. Fredrickson, T.-A. Roberts, Objectification theory: Toward understanding women's lived experiences and mental health risks, Psychology of Women Quarterly 21 (1997) 173–206. URL: https://doi.org/10.1111/j.1471-6402.1997.tb00108.x. doi:10.1111/j.1471-6402.1997.tb00108.x.

[2] J. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, Computers in Human Behavior 52 (2015) 436–442. URL: https://doi.org/10.1016/j.chb.2015.06.024. doi:10.1016/j.chb.2015.06.024.

[3] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sEXism Identification in Social neTworks, Procesamiento del Lenguaje Natural 67 (2021) 195–207. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

[4] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: sEXism Identification in Social neTworks, Procesamiento del Lenguaje Natural 69 (2022).

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All You Need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[6] 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.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[9] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in Spanish and English tweets using monolingual and multilingual BERT and ensemble models, CoRR abs/2111.04551 (2021). URL: https://arxiv.org/abs/2111.04551. arXiv:2111.04551.

[10] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. A. Smith, Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, CoRR abs/2002.06305 (2020). URL: https://arxiv.org/abs/2002.06305. arXiv:2002.06305.

[11] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, CoRR abs/2111.09453 (2021). URL: https://arxiv.org/abs/2111.09453. arXiv:2111.09453.

[12] D. Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 9–14. URL: https://aclanthology.org/2020.emnlp-demos.2. doi:10.18653/v1/2020.emnlp-demos.2.

[13] X. Han, J. Eisenstein, Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english, CoRR abs/1904.02817 (2019). URL: http://arxiv.org/abs/1904.02817. arXiv:1904.02817.

[14] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, Processing 150 (2009).

[15] TwitterSentimentDataset: Dataset de sentimientos en espanol, https://github.com/garnachod/TwitterSentimentDataset, 2015.

[16] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines, CoRR abs/2006.04884 (2020). URL: https://arxiv.org/abs/2006.04884. arXiv:2006.04884.

[17] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, Y. Artzi, Revisiting few-sample BERT fine-tuning, CoRR abs/2006.05987 (2020). URL: https://arxiv.org/abs/2006.05987. arXiv:2006.05987.

[18] M. Guzman-Silverio, Á. Balderas-Paredes, A. P. López-Monroy, 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, volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 293–302. URL: http://ceur-ws.org/Vol-2664/mexa3t_paper9.pdf.

[19] V. Gómez-Espinosa, V. Muñiz-Sánchez, A. P. López-Monroy, Transformers pipeline for offensiveness detection in mexican spanish social media, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, CEUR-WS.org, 2021.