EXIST2021: Detecting Sexism with Transformers and Translation-Augmented Data

Guillem García Subies[1]

Instituto de Ingeniería del Conocimiento, Francisco Tomás y Valiente st., 11 EPS, B
Building, 5th florr UAM Cantoblanco. 28049 Madrid, Spain
guillem.garcia@iic.uam.es

Abstract. This paper describes a system created for the EXIST 2021 shared task, framed within the IberLEF 2021 workshop. We present an approach mainly based in fine-tuned BERT models and Data Augmentation with translation and backtranslation. We show an approach to face multilingual problems augmenting the data without the overfitting that an aggressive backtranslation can generate. Our models far outperform the baselines and achieve results close to the state-of-the-art.

Keywords: Sexism Detection · BERT · Transformers · Data Augmentation · Backtranslation · Multilingual Corpora

1 Introduction

With the crescent trends in social rights and equal rights demands, NLP can help in detecting harmful and sexist behaviors. The EXIST (sEXism Identification in Social neTworks) [16] shared task proposes, during this third edition of the IberLEF [11] workshop, a dataset to detect sexism in it's most broad definition and also kinds of sexism.

This article summarizes our participation in all the EXIST tasks. Given the success of Transformer-inspired language models [20], both in academia and industry [21], we decided to use already pre-trained BERT [4] models. Specifically, we face the multilingual problem using different models for every language. We also conjecture that a good way to augment the data in multilingual problems is translating the data into the other languages of the dataset, so instead of having n_i for every language, we have $\sum n_i$ samples for every language. As the dataset is not too big, we also explore the Data Augmentation with Backtranslation [17].

In the next section, we will briefly see some previous work related to this topic. In Section 3 we will go through a brief description of the tasks and the corpora. Then, in Section 4, we will explain the main ideas behind the proposed models. In Section 5 we will present a summary of the experiments we carried out and the results we got. Finally, in Section 6 we will expose the main conclusions of our work and results and we will also propose some ideas for future work.

IberLEF 2021, September 2021, Málaga, Spain.

2 Related Work

There is an extensive bibliography on Sentiment Analysis and text classification in social networks, however not that much work has been done about identifying and classifying sexist behaviors in different languages.

For instance, Anzovino et al. [1] propose a sexism classification dataset also in Spanish and in English and proposed some solutions based on n-grams and classic machine learning models like SVMs. Following that line, Frenda et al. [5] use the previous dataset (only the English part) in combination with others to detect both misogyny and sexism following a similar approach of classic NLP.

More recent work by Grosz et al. [7], focuses on the sexism in the workplace and they obtain stat-of-the art results with GloVe embeddings and modified LSTM so they have attention mechanisms.

Another recent corpus, created by Rodríguez-Sánchez et al. [15], is focused on the detection of a broad amount of sexist behaviors in Spanish tweets, from the most explicit abuses to some more subtle expressions. They also showed that BERT-based models perform better that classic approaches or bidirectional LSTMs.

3 Tasks Description

The main corpus consists of 6977 tweets both in English and Spanish for the train split and 3368 tweets and 982 “gabs” (from gab.com) also in both languages for the test one.

The first task, Sexism Identification, consists of classifying tweets between **sexist** and **non-sexist**. There are 3600 **non-sexist** tweets and 3377 **sexist tweets**, so we can consider that the problem is well-balanced. The metric used for this dataset is the Accuracy.

For the second task, Sexism Categorization, there are six classes: **non-sexist**, **ideological-inequality**, **stereotyping-dominance**, **misogyny-non-sexual-violence**, **sexual-violence** and **objectification**. As we can see in Table 1, the dataset is unbalanced, so the F-measure is used as the ranking metric. We can see a similar distribution of the classes if we split the corpora into Spanish and English.

Class		**Nº Samples Task1**	**Nº Samples Task2**
non-sexist		3600	3600
ideological-inequality	sexist	3377	866
stereotyping-dominance			809
misogyny-non-sexual-violence			685
sexual-violence			517
objectification			500

Table 1. Distribution of Samples

In the table below, we can see some illustrative examples of the data and their labels:

I love poetry books, so I'm reading the one i have on this plane flight and one of the flight attendants (black women) goes "it's good to see a brotha reading something that's is so deep"	non-sexist
Can the fellas participate or is this just for the ladies/Non binary people because I don't wanna get clowned.	ideological-inequality
ive been sooo interesting my whole life and i just want to be a boring trophy wife now	stereotyping-dominance
Fucking skank	misogyny-non-sexual-violence
Bitches be begging me to fw them just to give me a reason not to fw them. Lol	sexual-violence
some women just don't deserve onlyfans, bitches be UGLY as fuck and ask you to pay $20 to see their UGLY FAT BLOTCHED TITTIES, BITCH!	objectification

Table 2. Examples of the different classes

4 Models

4.1 Data Preprocessing

We performed a simple preprocessing where we substituted some expressions with a more normalized form:

- Every URL was replaced with the string "[URL]" so we don't get strange tokens when the tokenizer tries to process a URL. Furthermore, no semantic information about sexism can be inferred from a URL, the only information relevant for the model is that there is a URL in that token.
- The hashtag characters ("#") were deleted ("#example" → "example") because the base language models we will use, are trained in generic text and might not understand their meaning. Furthermore, most of hashtags are used the same way as normal words.
- We replaced every username with the string "[USER]" because the exact name of a user does not really add any information about the sexism. The only relevant feature is knowing if someone was mentioned or not, but not who.
- Finally we normalized every laugh ("jasjajajajj" → "haha") so we minimize the noise of the misspellings, common in social networks.

4.2 Baselines

We created some baselines so we can compare our models properly. We selected a HashingVectorizer + RandomForest and a multilingual BERT (mBERT). This way, we can compare our models to a classic feature extraction model and a simple BERT-based one.

4.3 Language Models

We decided to fine-tune one language model for every language. For the Spanish language, we selected BETO [3], a BERT model trained with the Spanish Unannotated Corpora (SUC) [2] that has proven to be much better than the multilingual BERT model. For the English part of the dataset we used the original BERT model [4].

For the second task, given the imbalance in the classes, we performed a hierarchical classification, where the model from Task 1 classifies between **sexist** and **non-sexist** and another model is trained to detect specific kinds of sexism.

In addition, for the fine-tuning process, we carried out a Grid-search optimization over the main parameters of the neural network: learning rate, batch size and dropout rate. The search was performed with a 5-fold stratified cross-validation with the following grid: Learning rate, $(1e-6, 1e-5, 3e-5, 5e-5, 1e-4)$; batch size, $(8, 16, 32)$ and dropout rate, $(0.08, 0.1, 0.12)$. The best parameters for both models were: learning rate, $1e-5$; batch size, 16 and dropout rate, 0.1.

4.4 Data Augmentation

As the dataset is relatively small, we decided to run Data Augmentation techniques. We followed two different strategies to increase the amount of data in the corpora; Backtranslation [17] and translation of the different languages in the dataset.

Backtranslation This method consists of translating the samples into a pivot language and then translating them back into the original language. Given that the existing translation methods are not perfect, we get samples that are written in a slightly different way, but keep the original meaning. In particular, this technique has been proven useful for sentiment analysis and with twitter corpora before [10]. In this case, we used 30 pivot languages. For the translations we used the translation models of Helsinki NLP [19] based on the Marian model [9] and the Google Translate API for the ones that were not available in the Helsinki NLP models.

The selected languages are the following (expressed in ISO 639-1): *eu*, *la*, *zh-cn*, *hi*, *bn*, *pt*, *ru*, *ja*, *pa*, *mr*, *te*, *tr*, *ko*, *fr*, *de*, *vi*, *ta*, *ur*, *it*, *ar*, *fa*, *ha*, *kn*, *id*, *pl*, *uk*, *ro*, *eo*, *sv* and *el*. Also *es* and *en* were used for English and Spanish datasets respectively.

For every sample in the corpus, we randomly picked one pivot language to perform the backtranslation, so we ended up with a corpus of twice the size.

Multilingual Translation Following the above reasoning, we can also use labeled data (with the same gold standard) in other languages. So we translated every English sample into Spanish and *vice-versa*. This way, we should have a more robust training and avoid overfitting because the "new" samples are completely new for that language's model, opposed to the slightly modified samples from Backtranslation.

5 Experiments and Results

5.1 Experimental Setup

We trained all the models with a NVIDIA Tesla P100-PCIE-16GB GPU and a Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz CPU with 500GB of RAM memory.

The software we used was Python3.8, transformers 4.5.1 [21], pytorch 1.8.1 [13] and scikit-learn 0.24.1 [14].

5.2 Results

In the Table 3 we can see the results for our models in the test set of the first task. Note that the Tfid+SVM baseline and the AI-UPV_team (winner of the task) are taken from the task Overview [16]. The runs we presented for the contest were *2BERTs+Backtranslation*, *2BERTs* and *2BERTs+Multilingual translation* where 2BERTs refer to the different models used for each language, explained in Section 4.3. Note that our results correspond to the *GuillemGSubies* team in the official leaderboard.

The Backtranslation models had good results in our first training experiments, however they proved to overfit a lot for this task with an accuracy of 0.7479, just a bit better than the multilingual BERT baseline (0.734), but worse than just fine-tuned BERTs. This shows that Data Augmentation techniques are not always useful. Next, we can see that the Multilingual Translation models obtained an accuracy of 0.7683, which proves a better generalization than the model without any augmentation (0.7603). With this, our model is positioned very close to the best result in the competition, that is only 1.58% better.

Model	**Accuracy**
HV+RF	0.6830
Tfidf+SVM	0.6845
mBERT	0.7341
2BERTs+Backtranslation	0.7479
2BERTs	0.7603
2BERTs+Multilingual translation	0.7683
AI-UPV_team	0.7804

Table 3. Results for task1

For the second task, the results were similar to the ones obtained in the first task. In the Table 4 we can look at them in more detail. Again, the Tfidf+SVM baseline and the AI-UPV_team results come from the task Overview [16]. We can see that our models behaved consistently like in the first task, but the results were not that good. Despite that, the results are still very close to the best.

Model	**F-measure**
Tfidf+SVM	0.3950
HV+RF	0.4131
mBERT	0.4961
2BERTs+Backtranslation	0.5174
2BERTs	0.5218
2BERTs+Multilingual translation	0.5295
AI-UPV_team	0.5787

Table 4. Results for task2

To sum up the improvements of our models, we can see an ablation study for our best model (*2BERTs+Multilingual translation*) in the task 1 where each entry has a feature removed from the best model. This proves that most of the ideas introduced, produced some kind of improvement to the system. The most significant improvement was the selection of good hyperparameters for the model. Finally, it is also very remarkable that we get a large improvement by Multilingual Translation, proving our hypothesis about the ability of this Data Augmentation technique to generalize in Multilingual corpora.

Model	**Accuracy**
Best model	0.7683
Default model (no Grid-Search)	0.7451
Uncased	0.7599
No augmentation	0.7603
No preprocessing	0.7678

Table 5. Ablation study for the task1 models

6 Conclusions and Future Work

Through this shared task, we have seen that NLP can be of great help in detecting and classifying unwanted toxic and sexist behavior in social networks and there is still a long way to go.

The results obtained by our systems are very promising given their great performance and their simplicity. Furthermore, we proposed a new way of facing multilingual problems that provides better results. All this is very significant and could lead to much better results when combined with other improvements from the state-of-the-art.

We believe that our results could improve a lot using specific language models trained with corpora from social networks like TWilBert [6] for Spanish and BERTweet [12] for English. Another interesting approach would be to use a general language model and further pre-train it with corpora from the same domain [18] as the final task. These corpora would be easy to obtain given that the authors of the EXIST2021 shared task, gathered it from a list of keywords [16]. Finally, we have proven that good hyperparameters are also key for a good neural network so a better search, like the Population Based Training [8], would further improve the model.

Acknowledgments

This work has been partially funded by the Instituto de Ingeniería del Conocimiento (IIC) and the hardware used was also provided by the IIC.

References

1. Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on twitter. In: NLDB (2018)
2. Cañete, J.: Compilation of large spanish unannotated corpora (May 2019). https://doi.org/10.5281/zenodo.3247731, https://doi.org/10.5281/zenodo.3247731
3. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: to appear in PML4DC at ICLR 2020 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
5. Frenda, S., Ghanem, B., Montes-y Gómez, M., Rosso, P.: Online hate speech against women: Automatic identification of misogyny and sexism on twitter. Journal of Intelligent & Fuzzy Systems **36**(5), 4743–4752 (2019)
6. Ángel González, J., Hurtado, L.F., Pla, F.: Twilbert: Pre-trained deep bidirectional transformers for spanish twitter. Neurocomputing (2020). https://doi.org/https://doi.org/10.1016/j.neucom.2020.09.078, http://www.sciencedirect.com/science/article/pii/S0925231220316180
7. Grosz, D., Conde-Cespedes, P.: Automatic detection of sexist statements commonly used at the workplace (2020)
8. Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W.M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., Fernando, C., Kavukcuoglu, K.: Population based training of neural networks (2017)
9. Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-4020, https://www.aclweb.org/anthology/P18-4020

10. Luque, F.M.: Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis (2019)
11. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Ángel Álvarez Carmona, M., Álvarez Mellado, E., de Albornoz, J.C., Chiruzzo, L., Freitas, L., Adorno, H.G., Gutiérrez, Y., Zafra, S.M.J., Lima, S., de Arco, F.M.P., Taulé, M.: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). In: CEUR Workshop Proceedings (2021)
12. Nguyen, D.Q., Vu, T., Nguyen, A.T.: BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2020)
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L.: Automatic classification of sexism in social networks: An empirical study on twitter data. IEEE Access **8**, 219563–219576 (2020)
16. Rodríguez-Sánchez, F., de Albornoz, J.C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural **67**(0) (2021)
17. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 86–96. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1009, https://www.aclweb.org/anthology/P16-1009
18. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? (2020)
19. Tiedemann, J., Thottingal, S.: OPUS-MT — Building open translation services for the World. In: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT). Lisbon, Portugal (2020)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
21. 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., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: 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://www.aclweb.org/anthology/2020.emnlp-demos.6