EXIST 2021: Inducing Bias in Deep Learning Models

Ismael Garrido-Muñoz[1], Arturo Montejo-Ráez[1], and Fernando Martínez-Santiago[1]

CEATIC. Universidad de Jaén, España

Abstract. For the EXIST 2021 sexism detection task, a novel approach is proposed. In previous work we see that deep learning models capture biases related to race, gender, religion, etc. There are numerous works that seek to lower the bias of these models to make them fairer and more equitable. We will use the opposite approach to these works, instead of reducing the bias of the models we will seek to increase it. So we will train 3 times the same base model, generating 3 versions of the model. A normal version, a sexist version and an anti-sexist version. Belonging to these models will mean marking or not the tweet as sexist.

Keywords: BERT · fine-tune · bias.

1 Introduction

The EXIST 2021 workshop deals with the identification and classification of sexist messages coming from the websites gab.com and twitter.com. Sexism detection is not a novel task, we could consider it a specialization of the hate detection task, in which the hate has an impact on one of the genders. Although it is not a novel task, it is a topical task. There are already systems in production that use hate detection as an automatic content moderation tool, for example twitter [14] has recently started to warn its users when it detects a hateful message before they are sent, asking the user to review them. It is not the only social network that is working on detecting this type of behavior, we can find similar initiatives in Facebook [13] or even Google. In the case of Google we can see how despite the good intentions, its AI seems to be race biased [11].

This is not a trivial task since in many cases it is the social context itself that causes a message to undervalue the role of one of the genders or even feed negative stereotypes. This document explains the approach followed for the participation in the first version of the EXIST 2021 workshop along with the data and tools used.

IberLEF 2021, September 2021, Málaga, Spain.

1.1 EXIST Task

EXIST is proposed as a workshop embedded as part of IberLEF. It provides a set of tweets and gabs in Spanish and English and proposes two tasks. The first one is to detect whether each of the messages is sexist or not. The second task proposes to classify the type of sexism present in the tweet for a set of proposed categories.

The task considers that a message is sexist as one that produces prejudice, inequality or stereotype following The Oxford English Dictionary definition. They also provide a set of tweets and gabs labeled in the proposed tasks.

1.2 Previous\Related Work

There are many previous works [12,2] on the task of hate speech detection but we will focus on deep learning approach and on treating sexism as a bias. Studies such as [9] approaches the task as a problem of racial bias in the datasets. In [8] they fine-tune BERT by adding additional layers and the best result is obtained by adding a CNN layer. [3] use three deep-learning architectures (CNN, LSTMs and FastText) to classify tweets as racist, sexist or neither.

2 Approach

We will try a novel approach, instead of using a classifier we will train three models of each language (English and Spanish) and we will score how likely are the tweets to belong to each model. One model base model will be trained with random data, the stereotyped model will be trained with toxic/stereotyped comments and the anti-stereotype will be trained with data sample that contradicts stereotypes. Given a sentence, it will be considered toxic/sexist if the likelihood of been generated from a stereotyped model is greater than that of any of the other two models (anti-stereotyped and neutral).

2.1 Building the models

Previous work demonstrates that neural network based models capture certain associations present in the data that may be sexist, racist, or have some type of unwanted bias. The study of bias in neural network based models shows how recent models such as BERT or even GPT-3 capture these biases. Studies such as [7] show that from version two to version three of GPT model, it does substantially improve the ability to generalize and emulate in language. However, it also captures negative race-related biases [1] better than its predecessor or even improves the model's ability to generate extremist text.

One of the most common examples is that Word embeddings models captures the association man - woman → king - queen correctly, however it also captures associations like man - woman → computer scientist - homemaker

Using a list of professions it is obvious that there are professions strongly associated with men and others strongly associated with women. Despite the fact that in certain applications this bias may be detrimental, we will take it as an advantage.

2.2 Resources

Civil Comments/Jigsaw corpora This dataset is the one proposed by Google's Jigsaw initiative for the [5] "Jigsaw Unintended Bias in Toxicity Classification" challenge on the Kaggle platform. The data are comments released by the Civil Comments platform, tagged by humans to categorize them as toxic, obscene, threatening, insulting, or identity hate. In addition, tags are included to indicate what race, gender, sexual orientation, religion, ethnicity or disability is addressed in the message. For the detection of sexism, gender labels with high toxicity values are the ones we consider most important.

StereoSet corpora It is a dataset in English labeled to reflect the stereotypes present in the data. To do this, a dataset is built[10] in which spaces are left in the sentences to be filled in with some various words. For each of these phrases, we have a label indicating what type of stereotype is being tested (race, gender, ...) and another label indicating the result from the review of 5 people indicating whether the phrase reflects a stereotype, an anti-stereotype or if the phrase simply does not make sense or is not related to the studied stereotype.

English BERT model [6] BERT is a general-purpose language model trained on a large corpus. The difference between BERT and other models is that the context of each word is formed by both left-hand and right-hand terms of the same word throughout the network.

Multiple versions of BERT are available, we have used bert-base-uncased.

Spanish BERT model (BETO) [4] BETO is equivalent to BERT for Spanish. It is trained by the same technique as BERT (Whole Word Masking) using the Spanish Unannotated Corpora dataset. Among these data we can find 300 million lines in Spanish from different datasets, including wikis, europarliament texts, united nation texts, open subtitles, among many others. We will use bert-base-spanish-wwm-uncased version.

2.3 Other resources

For working with the models the HuggingFace libraries were used. Some scripting and data transformation work was done using custom nodejs/python scripts.

2.4 Training process

We have taken the English and Spanish models described above and fine-tuned them using the two stereotyped corpora introduced before. The fine-tune task was the same as the common task used to train these two models from scratch: masked language modeling. Training process was performed on a machine equipped with x2 Intel(R) Xeon(R) Silver 4208 CPUs at 2.10GHz 192 GB of RAM and x2 NVIDIA Tesla V100 (32GB, Volta). The implementation was based on the PyTorch versions of these models provided by the Transformers library. It took about 10 minutes for training the Spanish version of the three models (stereotype, anti-stereotyped and neutral) and about 20 minutes for the three English ones.

2.5 Likelihood

We calculate the probability of a sentence belonging to the model by using the chain rule of probability. To get the normalized probabilities we apply the softmax function to the model output logits which are the raw predictions of the last layer of the network.

$$\begin{gathered} P(X_1, X_2, \cdots, X_n) = \\ P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdots P(X_n|X_{n-1}, X_{n-2}, \cdots, X_1) = \\ \textstyle\prod_{k=1}^{n} P(X_k|X_1^{k-1}) \end{gathered} \tag{1}$$

https://huggingface.co

3 Results and discussion

It has been verified with some sentences from the training set that the model does indeed capture some structures in the form of probabilities. For example for the sentence "The schoolgirl is walking down the street, She is a target for child predators", the probabilities are:

	AntiStereotype	Other	Stereotype
the	0.00894868653267622	0.032950423657894135	0.005774959456175566
schoolgirl	0.0002689003486011643	0.00017878220387501642	0.002795979264192283
is	0.006179201416671276	0.03133246302604675	0.00566072016954422
walking	6.194096204126254e-05	3.609031409723684e-05	0.0002682682534214109
down	9.467829659115523e-05	2.7176944058737718e-05	6.732833571732044e-05
the	0.026873258873820305	0.02360343001782894	0.023960916325449944
street	0.005731010489128607	0.0007263075825676424	0.1331725064192142
she	0.001979067223146558	0.0008687127847224474	0.0036194631829857826
is	0.0073087140917778015	0.02685435861349106	0.08509364724159241
a	0.0025236799847334623	0.0009333917987532914	0.0032958367373794317
target	1.7575850506545976e-05	1.2269218530036596e-07	1.707993391164564e-07
for	0.00017388597188983113	3.896695488947444e-05	8.425339183304459e-05
child	0.022296445444226265	0.0001155364152509719	0.011923404410481453
predators	0.004697749387711762	0.0002539085783155315	0.001255747340716577
Total	1.5692968748328783e-40	9.478321051397374e-48	1.1521866603985196e-39

This phrase is effectively classified as stereotypical. However, the difference with respect to anti-stereotyped, which is the opposite class, is not large enough. This may be due to the fact that the probability calculation method does not treat BERT as bidirectional.

4 Conclusions and future work

We have presented a proposal for the classification of sexist tweets based on bias Although we see that there are differences between the models, there is still a lot of work to be done for this approach to give good results. On the one hand, it will be necessary to find concrete structures that are representative of sexist or toxic content. Once the structures have been found, it will be necessary to improve the training method to emphasize the structures and thus fine-tune the model by focusing on the characteristics of the problem.

Another point where improvement can be made is in the calculation of model membership, and other metrics can be applied to obtain a more reliable result.

One aspect that would be interesting to explore is the automatic extraction of social media content to train the model. Is it feasible to characterize as toxic or non-toxic non-annotated texts from social networks using these networks' own metrics? For example, characterizing as toxic those with a large proportion of negative votes and as non-toxic those with a proportionally large number of positive votes. This can allow the model to mature, having the same information that fits the reality.

References

1. Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models. CoRR **abs/2101.05783** (2021), `https://arxiv.org/abs/2101.05783`
2. Al-Hassan, A., Al-Dossari, H.: Detection of hate speech in social networks: a survey on multilingual corpus. In: 6th International Conference on Computer Science and Information Technology. vol. 10 (2019)
3. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. p. 759–760. WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3041021.3054223, `https://doi.org/10.1145/3041021.3054223`
4. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pretrained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
5. Conversation AI, J.: Jigsaw unintended bias in toxicity classification — kaggle (2021), `https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview`
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), `http://arxiv.org/abs/1810.04805`
7. McGuffie, K., Newhouse, A.: The Radicalization Risks of GPT-3 and Advanced Neural Language Models. arXiv e-prints arXiv:2009.06807 (Sep 2020)
8. Mozafari, M., Farahbakhsh, R., Crespi, N.: A bert-based transfer learning approach for hate speech detection in online social media. CoRR **abs/1910.12574** (2019), `http://arxiv.org/abs/1910.12574`
9. Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on bert model. PLOS ONE **15**(8), 1–26 (08 2020). https://doi.org/10.1371/journal.pone.0237861, `https://doi.org/10.1371/journal.pone.0237861`
10. Nadeem, M., Bethke, A., Reddy, S.: Stereoset: Measuring stereotypical bias in pretrained language models. CoRR **abs/2004.09456** (2020), `https://arxiv.org/abs/2004.09456`
11. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1668–1678. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1163, `https://www.aclweb.org/anthology/P19-1163`
12. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media. pp. 1–10 (2017)
13. Schroepfer, M.: How ai is getting better at detecting hate speech (Nov 2020), `https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech`
14. Statt, N.: Twitter tests a warning message that tells users to rethink offensive replies (May 2020), `https://www.theverge.com/2020/5/5/21248201/twitter-reply-warning-harmful-language-revise-tweet-moderation`