2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService) | 978-1-6654-5890-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigDataService55688.2022.00036

Fine-Tuning Language Models to Mitigate Gender Bias in Sentence Encoders

Tommaso Dolci
Politecnico di Milano – DEIB
Milan, Italy
tommaso.dolci@polimi.it

***Abstract*—Language models are used for a variety of downstream applications, such as improving web search results or parsing CVs to identify the best candidate for a job position. At the same time, concern is growing around word and sentence embeddings, popular language models that have been shown to exhibit large amount of social bias. In this work, by leveraging the possibility to further train state-of-the-art pre-trained embedding models, we propose to mitigate gender bias by fine-tuning sentence encoders on a semantic similarity task built around gender stereotype sentences and corresponding gender-swapped anti-stereotypes, in order to enforce similarity between the two categories. We test our intuition on two popular language models, BERT-Base and DistilBERT, and measure the amount of gender bias mitigation using the Sentence Encoder Association Test (SEAT). Our solution shows promising results despite using a small amount of training data, proving that post-processing bias mitigation techniques based on fine-tuning can effectively reduce gender bias in sentence encoders.**

***Index Terms*—Natural language processing, gender bias, word embeddings**

I. Introduction

Natural language processing (NLP) is becoming an integral part of our everyday life, including applications such as conversational agents and web search suggestions. Recently, pre-trained word embedding models such as BERT [1] provided extremely flexible and powerful text representations, to be used for a variety of natural language understanding tasks and downstream applications, from CV parsing to hate-speech detection. Word embeddings are a popular method to represent words as vectors, so that the geometry between vectors captures the semantic relationships between the corresponding word. Similarly, it is possible to retrieve sentence vectors as representations of entire sentences. However, despite the enthusiasm from NLP researchers, word embeddings have been shown to learn and exhibit common stereotypes of the Western society [2], [3]. As a consequence, the social bias in the geometry of the model is reflected in downstream applications; for instance, gender-biased models tend to associate computer science to men and not women, thus preferring men for computer-related jobs, or giving more visibility to male researchers on search engines [2]. Over the recent years, a variety of benchmarks and tests have been designed to detect and quantify social bias in word embeddings, regarding attributes such as gender, ethnicity and religion [4]–[6]. Some bias mitigation techniques have been proposed, mainly focusing on simple word embedding models [2], [7].

In this preliminary work, our goal is to explore a post-processing technique to mitigate gender bias in pre-trained language models, by fine-tuning them on a semantic textual similarity task. The general idea is to re-balance gender disparity by enforcing similarity between pairs of sentences, the first containing a gender stereotype, and the second containing the corresponding gender-swapped anti-stereotype. We focus on two popular word and sentence embedding encoders, namely BERT-Base [1] and DistilBERT [8], part of the BERT family of Transformer-based language models. We test the resulting fine-tuned models on the Sentence Encoder Association Test (SEAT) [4], a popular framework to discover implicit stereotypical associations in language encoders. Our solution contributes to the research field of bias and fairness in NLP, paving the way for gender bias mitigation techniques based on fine-tuning, focusing on language models for which re-training from scratch is extremely expensive and frequently unfeasible.

II. Methodology

The first step of our research consists in further training pre-trained sentence encoders on a semantic textual similarity task. To do so, we employ SentenceTransformers[1] a framework that offers a large collection of language models and allows to fine-tune them on a variety of NLP tasks. In particular, we focus on a custom-made task based on semantic similarity evaluation between pairs of sentences. In fact, semantic similarity tasks have the advantage of being generic enough for the resulting model to be still used for a variety of downstream applications. Since gender bias in language models is largely caused by the internalisation of gender stereotypical conceptions, we build a training dataset consisting of pairs of sentences containing gender stereotypes and anti-stereotypes. The core idea is to re-balance gender disparity previously learned by the model, by forcibly reducing the distance between the embeddings of the two sentences of the pair, that contain respectively female and male subjects. In other words, we want to tell the model that there is no correlation between gender and these common stereotypes. Since stereotypical situations are expressed by the context of the sentence, this approach is particularly significant for models based on contextualized word embeddings, such as those of the BERT family, on which we focus.

[1] https://www.sbert.net/index.html

DOI 10.1109/BigDataService55688.2022.00036

TABLE I
EXAMPLES OF SENTENCE PAIRS.

Stereotype	Anti-stereotype
She does not do any work.	He does not do any work.
A mother is caring.	A father is caring.
She is in the kitchen cooking.	He is in the kitchen cooking.

To build the input data for training, we select sentences from StereoSet [5] and CrowsPairs [9], two collections of social stereotypes in text format. We consider gender stereotypes only, and perform a duplication on each entry by swapping all gender words to their female/male counterpart. This way, we obtain pairs of sentences containing respectively a gender stereotype and its corresponding anti-stereotype with subjects of the opposite gender, as illustrated in Table I. The resulting dataset is relatively small, comprising 848 pairs of sentences split into training, validation and testing sets, with a ratio of 80/10/10 respectively. Each sentence pair is annotated with a normalized gold similarity score, which for this task we want to be the highest possible to force the network into equalizing the two genders. Therefore, each pair is assigned a gold similarity score equal to 1. To compute loss during training, we use a function based on cosine similarity, a popular similarity metric for word and sentence embeddings.

After fine-tuning, we want to evaluate the resulting models and measure the amount of gender bias mitigation. To do this, we select gender-related tests from SEAT [4], i.e. tests C6, C6b, C7, C7b, C8 and C8b. SEAT measures the association between two sets of targets and two sets of attributes. For instance, test C6 has male and female names as targets, and sentences regarding career and family as attributes. The magnitude of the association is measured by the effect size, based on the cosine similarities between embeddings of the target concepts and embeddings of attributes: the higher the effect size, the stronger is the stereotypical association between targets and attributes. More details and the official code for running SEAT are available on the authors' GitHub page[2].

III. Experimental Results

We test our intuition on two language models of the BERT family, namely BERT-Base [1] and DistilBERT [8], a smaller, distilled version of BERT. In fact, language models based on BERT fall into the category of pre-trained models, and therefore it is possible to fine-tune them with minimal effort and computational resources. Additionally, both are popular models for a number of downstream tasks, where social bias is more dangerous. Results are shown in Table II. Both models have already been further trained on a natural language inference task. We add 4 epochs of fine-tuning on the semantic similarity task described in the previous section. For both BERT-Base and DistilBERT the effect size decreases in almost all tests, especially in tests C6 and C6b. The reduction in the average score is also significant, particularly in the case of BERT-Base, where all tests demonstrate a much lower degree of stereotypical association.

[2] https://github.com/W4ngatang/sent-bias

TABLE II
EFFECT SIZE FOR GENDER-RELATED TESTS IN SEAT—COMPARISON BETWEEN THE ORIGINAL BERT-BASE AND DISTILBERT, AND THEIR RESPECTIVE FINE-TUNED (FT) MODELS. BEST RESULTS UNDERLINED.

Test	BERT	BERT-FT	Distil	Distil-FT
C6: Career/Family	1.580	0.813	1.442	0.562
C6b: Career/Family	0.794	-0.004	0.708	0.135
C7: Math/Arts	0.529	0.412	0.732	0.756
C7b: Math/Arts	0.327	0.040	0.627	0.474
C8: Science/Arts	0.594	0.160	0.986	0.844
C8b: Science/Arts	0.763	0.273	1.141	1.312
Avg Abs Effect Size	0.765	0.284	0.939	0.681

IV. Conclusions and Future Work

In this paper, we showed the possibility to mitigate gender bias by fine-tuning sentence encoders on a semantic similarity task built around gender stereotype sentences and corresponding gender-swapped anti-stereotypes. Despite using a small amount of training data, results are very promising, proving that such post-processing bias mitigation techniques can effectively reduce gender bias in sentence encoders.

Future work includes experimenting with different language models and improving the data used for the fine-tuning procedure. In fact, not considering the small amount of sentences employed, both StereoSet and CrowsPairs have been criticized for their lack of quality and a number of pitfalls [10]. Furthermore, it is important to evaluate the amount of mitigation with different tests, such as WinoBias [6]. In fact, despite its popularity, SEAT frequently gives mixed and unexpected results for contextualized word embeddings. Finally, a thorough review and comparison with state-of-the-art studies on bias mitigation in language models is also fundamental.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[3] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 2019, pp. 166–172.

[4] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," *arXiv preprint arXiv:1903.10561*, 2019.

[5] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[6] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.

[7] S. Bordia and S. R. Bowman, "Identifying and reducing gender bias in word-level language models," *NAACL HLT 2019*, p. 7, 2019.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[9] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," *arXiv preprint arXiv:2010.00133*, 2020.

[10] S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach, "Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets," in *ACL*, 2021.