DEBIASING THE PRE-TRAINED LANGUAGE MODEL THROUGH FINE-TUNING THE DOWNSTREAM TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have revealed that the widely-used pre-trained language models propagate societal biases from the large unmoderated pre-training corpora. Existing solutions mostly focused on debiasing the pre-training corpora or embedding models. Thus, these approaches need a separate pre-training process and extra training datasets which are resource-intensive and costly. Indeed, studies showed that these approaches hurt the models' performance on downstream tasks. In this study, we focus on gender debiasing and propose *Gender-tuning* which comprises of the two training processes: gender-word perturbation and fine-tuning. This combination aims to interrupt gender word association with other words in training examples and classifies the perturbed example according to the ground-truth label. Gender-tuning uses a joint-loss for training both the perturbation model and fine-tuning. Comprehensive experiments show that Gender-tuning effectively reduces gender biases scores in pre-trained language models and, at the same time, improves performance on downstream tasks. Gender-tuning is applicable as a plug-and-play debiasing tool for pre-trained language models. The source code and pre-trained models will be available on the author's GitHub page.

1 INTRODUCTION

In recent years, pre-trained language models have achieved state-of-the-art performance across various downstream tasks in natural language processing (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020). One of the crucial reasons for this success is pre-training from large-scale corpora, which are collected from unmoderated sources such as the internet. Prior studies have shown that pre-trained language models capture a significant amount of social biases existing in the pre-training corpus (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018; May et al., 2019; Kurita et al., 2019; Gehman et al., 2020). For instance, they showed that the pre-trained language models learn the word "he" is closer to the word "engineer" because of the high frequency of the co-occurrence of this combination in the training corpora, which is known as social biases. Since pre-trained language models are increasingly deployed in real-world scenarios, there is a serious concern that these language models propagate discriminative prediction and unfairness.

Several studies have focused on developing solutions for mitigating the social biases, including but not limited to using banned word lists (Raffel et al., 2020), building training datasets with more care and deliberation (Bender et al., 2021), balancing the biased and unbiased terms in the training dataset (Dixon et al., 2018; Bordia & Bowman, 2019), debiasing embedding spaces (Liang et al., 2020; Cheng et al., 2021), and self-debiasing in text generation (Schick et al., 2021). Although all these solutions have shown different levels of success, Meade et al. (2021) illustrated that the current debiasing techniques worsen language models' ability in various functionalities. For example, the solutions related to banned words prevent language models from gaining knowledge of topics related to banned words. Also, the current debiasing techniques hurt the pre-trained language model performance on downstream tasks (Meade et al., 2021). Furthermore, dataset

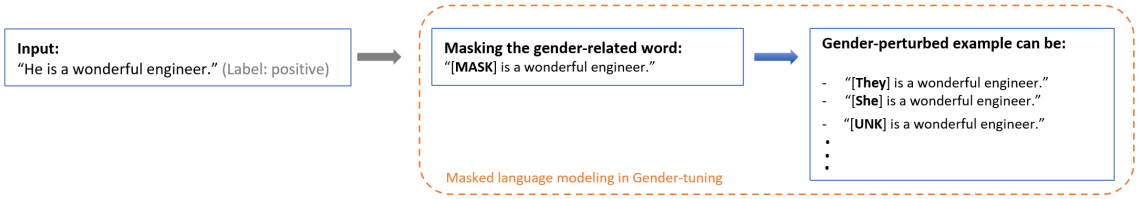

Figure 1: The illustration of what is happening for a training example through the first training step of Gender-tuning, masked language modeling. Gender-tuning masks the existing gender-related word(s) and predicts the masked word(s) with the word(s), which minimize masked language modeling training loss.

curation/augmentation and pre-training (two resource-intensive tasks) are needed for most of the above solutions (Schick et al., 2021).

To break social biases, we propose a debiasing method named *Gender-tuning* that comprises two subsequent training phases: bias perturbation and fine-tuning. The proposed method uses masked language modeling for the gender bias perturbation training phase. Gender-tuning masks the gender word(s) in an example and predicts the highest probable token (Figure 1) that minimizes the masking model loss. Then Gender-tuning classifies the gender-perturbed example through fine-tuning (second training phase) based on the ground-truth label and computes fine-tuning loss. Afterward, the proposed method uses an aggregation of losses generated from the two training phases called joint-loss. The joint-loss allows Gender-tuning to interrupt the association between the gender words and other words in training examples while preserving the ground-truth label. As a result, Gender-tuning trains avoid propagating the biases in the pre-trained language models when training on the downstream task and mitigate the social biases score in the pre-trained language models.

The key advantage of our method is integrating debiasing approach into fine-tuning setting. This allows the learning process to be carried out without requiring a separate pre-training or additional training data other than the downstream task dataset. Integrating with fine-tuning also makes Gender-tuning a plug-and-play debiasing tool for any pre-trained language models. We conducted comprehensive experiments following two state-of-the-art studies, sentence-based embedding debiasing (Sent-D) (Liang et al., 2020) and FairFil (FairF) (Cheng et al., 2021) to evaluate the effectiveness of the Gender-tuning. The results show that Gender-tuning reduces the gender biases in the pre-trained language model more accurately than in those studies while improving the downstream task performance. Furthermore, we reported the performance of the Gender-tuning using RoBERTa, which has BERT-based architecture with larger pre-training corpora and training steps. The results in both models (BERT and RoBERTa) proved that gender-tuning successfully reduces the gender bias scores in pre-trained language models. Indeed, our ablation study show that joint-loss training plays an essential role in Gender-tuning's success.

2 METHODOLOGY

In this section, we formally introduce the proposed approach setup named *Gender-tuning* and the insights behind it (Figure 2).

2.1 GENDER-TUNING

As shown in Figure 2, Gender-tuning develops the capabilities of fine-tuning's training process to alleviate the problem of social biases propagating when training on downstream task datasets and reducing the social biases

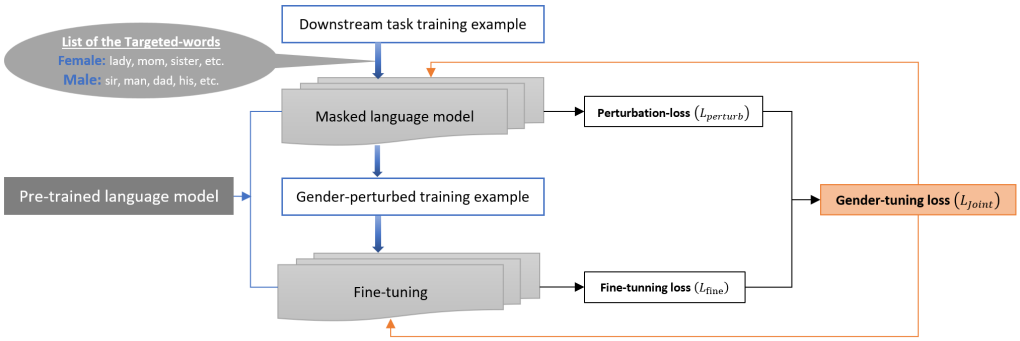

Figure 2: The proposed method, *Gender-tuning*, empowers fine-tuning to debias pre-trained language model by using a masked language model to perturb the linguistic relation between the gender-word(s) and other words in training examples and computes the perturbation training loss. Then Gender-tuning classifies the perturbed example based on the ground-truth label through fine-tuning and generates a fine-tuning loss. Finally, for fair training, Gender-tuning uses a joint-loss for training the masked language model and fine-tuning. The examples without any gender-word are directly fed to fine-tuning.

scores in the pre-trained language models. For this aim, Gender-tuning aggregate two training processes: 1) Gender-word(s) perturbation and 2) Fine-tuning.

Gender-tuning uses masked language modeling to perturb the relation between the gender words and other words in an example by masking the gender word(s) and predicting the word(s) which minimize the masked language model training loss and generates the *gender-perturbed* example. In this case, the final hidden vectors corresponding to the masked token(s) is fed into an output softmax over the embedding vocabulary same as a standard language model. If the i-th token(s) is chosen, Gender-tuning replace the i-th token(s) with the [MASK] token(s). Then the final hidden vector for i-th token(s) will be used to predict the masked token(s) with the aggregation of the cross-entropy loss from all masked token(s) that we denote as *perturbation loss* ($\mathcal{L}_{perturb}$) (Fig. 2).

Afterward, the gender-perturbed example created by the masked modeling's training process is fed into fine-tuning to be classified based on the ground-truth label (y). Then $p_\theta(y' = y|\hat{x})$ is the fine-tuning function to predict the gender-perturbed example's label (y') based on the gender-perturbed example ($\hat{x}$) and compute the fine-tuning training loss ($\mathcal{L}_{Fine}$), where θ is the pre-trained language model parameters for the fine-tuning. Finally, Gender-tuning will be trained based on a Gender-tuning loss ($\mathcal{L}_{Joint}$) that is a weighted aggregation of these two processes (i.e., masked modeling and fine-tuning):

$$\mathcal{L}_{Joint} = \alpha\ \mathcal{L}_{perturb} + \ (1-\alpha)\mathcal{L}_{Fine} \tag{1}$$

Where α is a weighting factor, we employ it to adjust the contribution of the two training losses in computing the Gender-tuning loss. The joint loss is used to train the masked language model and the fine-tuning in training iterations. For passing each training iteration, the training loss of both steps must be close to zero. Otherwise, the training continues until getting the smallest value for Gender-tuning training loss.

Combining the two training losses to compute the joint-loss helps the debiasing process in two ways. Firstly, suppose the masked language model creates an inconsistent example. For instance, the example: " the film affirms the power of the [actress]" changes to $\Rightarrow$ " the film affirms the power of the [science]", which is not only a non-related gender word but can change the concept of the example and raise perturbation loss value ($\mathcal{L}_{perturb.} > 0$). In this case, if fine-tuning classifies the perturbed example correctly and makes fine-loss

Table 1: The illustration of the different types of perturbation outputs such as neutral, same-gender, convert-gender, deleting, and identical that are generated by Gender-tuning.

Training input	Perturbed	Type	Label
with [**his**] usual intelligence and subtlety.	with [**the**] usual intelligence and subtlety.	neutral	1
by casting an [**actress**] whose face projects that [**woman**] 's doubts and yearnings , it succeeds.	by casting an [**image**] whose face projects that [**person**] 's doubts and yearnings , it succeeds.	neutral	1
certainly has a new career ahead of [**him**] if [**he**] so chooses.	certainly has a new career ahead of [**her**] if [**she**] so chooses.	convert-gender	1
by [**men**] of marginal intelligence , with reactionary ideas.	by [**people**] of marginal intelligence , with reactionary ideas.	neutral	0
why this distinguished [**actor**] would stoop so low.	why this distinguished [**man**] would stoop so low.	same-gender	0
it is very awful - - and oozing with creepy [**men**].	it is very awful - - and oozing with creepy [UNK] .	deleting	0
Proves once again [**he**] hasn't lost.	Proves once again [**he**] hasn't lost .	identical	1

close to zero, the aggregation of training losses from the perturbation and fine-tuning forces Gender-tuning to continue the training iteration and be updated.

Secondly, suppose Gender-tuning creates social gender bias through the process of gender perturbation. For instance, the example: "angry black [actor] changes to $\Rightarrow$ "angry black [woman]" that "woman" and "actor" are not close semantically and raise perturbation loss value ($\mathcal{L}_{perturb.} > 0$). In this case, the output of the fine-tuning might be correct ($\mathcal{L}_{fine} \approx 0$) based on the learned biases in the pre-trained language model. However, aggregation of two training losses generates a big join-loss that prevents fine-tuning from getting a reward for being correct and enforces the Gender-tuning to continue the training iteration and be updated.

2.2 PERTURBATION STRATEGY

The pre-trained language models achieved state-of-the-art performance on the downstream tasks datasets by applying the masked language model for the example perturbation in pre-training phase. Thus we hypothesize that the masked language modeling can generate realistic gender-perturbed examples that can considerably modify the gender relation between the input tokens without affecting the label. Furthermore, it is safe for consistency between two training phases of Gender-tuning, i.e., gender-words perturbation and fine-tuning. However, there is a concern that the pre-trained masked language model transfers the gender bias through the perturbation process.

For clarifying this concern, we investigate the predicted tokens that the pre-trained masked language model replaces with the gender-words. We randomly select 300 examples from training dataset including 150 examples with feminine words and 150 examples with masculine words. Based on these 300 examples, we observe five types of perturbation as shown through some examples in Table 1:

- **Neutral**; replace the gender-words with neutral word such as people, they, their, and etc.
- **Convert-gender**; replace the gender-words with opposite gender. the word "he" change to "she".
- **Same-gender**; replace the gender-words with the same gender. change the word "man" to "boy".
- **Deleting**; replace the gender-words with unknown token ([UNK]). In 300 examples, it only happens when there are several masked tokens.

- **Identical**; replace the gender word with itself. It mostly happens when there is only one gender word.

In our investigation with 300 examples, we had 46% Neutral, 29% Identical, 17% Convert-gender, 7% Same-gender, and 1% Deleting perturbation. As illustrated in Table 1, Gender-tuning does not make a meaningful change in identical and same-gender perturbation. These examples likely conform to the gender biases in the masked language modeling. Suppose identical, or same-gender perturbation gets the correct output from the perturbation process ($\mathcal{L}_{perturb.} \approx 0$). In this case, the only way to learn the biases in the masked language model is to get the correct output from fine-tuning step and joint-loss close to zero. This issue stops the masked and pre-trained language models from further update. However, joint-loss plays an essential role in alleviating learning gender bias from identical and same-gender perturbations.

To clarify the role of joint-loss in overcoming this problem, we investigated fine-tuning output on identical and same-gender perturbations. We observed that fine-tuning gets the incorrect output from 60% of the identical and 75% of the same-gender perturbation. Thus these examples return to training iteration because their joint-loss is big enough to update the language model and perform a new training iteration. New training iteration means re-perturbing and re-fine-tuning result on these examples. Therefore, training based on both training steps' loss and computing joint-loss persistently prevents learning from gender bias in masked modeling as well as the pre-trained language model.

3 EXPERIMENTS

We evaluate the effectiveness of Gender-tuning at reducing gender biases in pre-trained language models and its performance on downstream tasks. For comparison purposes, we follow two previous studies on embedding debiasing, Sentence debiasing (Sent-D) (Liang et al., 2020) and FairFilter (FairF) (Cheng et al., 2021).

3.1 DATASET

We conducted empirical studies on the following three tasks from the GLUE[1] benchmark (Wang et al., 2019): (1) **SST-2**: Stanford Sentiment Treebank is used for binary classification for sentences extracted from movie reviews (Socher et al., 2013). It contains 67K training sentences. (2) **CoLA**: Corpus of Linguistic Acceptability (Warstadt et al., 2019) consists of English acceptability judgment. CoLA contains almost 9K training examples. (3) **QNLI**: Question Natural Language Inference (Wang et al., 2018) is a QA dataset which is derived from the Stanford Question Answering Dataset (Rajpurkar et al., 2016) and used for binary classification. QNLI contains 108K training pairs. Also, we use the feminine and masculine word lists created by (Zhao et al., 2018) for gender-word perturbation in Gender-tuning.

3.2 BIAS EVALUATION METRIC

Following the prior studies, we use Sentence Encoder Association Test (SEAT) (May et al., 2019) to measure the gender bias scores in the pre-trained language models that trained using Gender-tuning. SEAT extended the Word Embedding Association Test (WEAT; caliskan2017semantics) to sentence-level representations. WEAT compares the distance of two sets. Two sets of target words (e.g., {*family, child, parent,...*} and {*work, office, profession,...*}) that characterize particular concepts $family$ and $career$ respectively. Two sets of attribute words (e.g., {*man, he, him,...*} and {*woman, she, her,...*}) that characterize a type of bias. WEAT evaluates whether the representations for words from one particular attribute word set tend to be more closely associated with the representations for words from one particular target word set. For instance, if the

[1]https://gluebenchmark.com/tasks

$female$ attribute words listed above tend to be more closely associated with the $family$ target words, this may indicate bias within the word representations.

Let's denote A and B as sets of attribute words and X and Y the set of target words. As described in (Caliskan et al., 2017) the WEAT test statistic is:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{2}$$

where for a specific word w , $s(w, A, B)$ is defined as the difference between w's mean cosine similarity with the words from A and w's mean cosine similarity with the word from B. They report an effective size given by:

$$d = \frac{\mu([s(x, A, B)]_{x \in X} - \mu([s(y, A, B)]_{y \in Y})}{\sigma([s(t, X, Y)]_{t \in A \cup B})} \tag{3}$$

where μ and σ denote the mean and standard deviation respectively. Hence, an effect size closer to zero represents smaller degree of bias in the word representation. The SEAT test extended WEAT by replacing the word with a collection of template sentences (i.e., *"this is a [word]", "that is a [word]"*). Then the WEAT test statistic can be computed on a given sets of sentences including attribute and target words using sentence representations from a language model.

3.3 Experimental setup

Two widely used pre-trained language models have been chosen for this study, BERT-base (Devlin et al., 2019)and RoBERTa-base (Liu et al., 2019). BERT-base is a bidirectional encoder with 12 layers and 110M parameters that is pre-trained on 16GB of text. RoBERTa-base has almost the same architecture as BERT but is pre-trained on ten times more data (160GB) with significantly more pre-training steps than BERT. We report the SEAT effect size for three different setups: (1) **Origin**: Directly fine-tuning the pre-trained model using huggingface (Wolf et al., 2020) on the three downstream task datasets. (2) **Gender-tuning**$_{random}$: Performing Gender-tuning with perturbing the input tokens randomly (5% of each input sequence) on the downstream task datasets. (3) **Gender-tuning** (our method): Performing Gender-tuning with perturbing the input tokens based on existing gender-word on the downstream task datasets. We use the same hyperparameter for all three setups for a fair comparison.

The hyperparameters of the models, except batch size, are set to their default[2] values (e.g., epoch=3, learning-rate = 2×10^{-5}, and etc.). After trying several trials run, the batch size has been selected among $\{8, 16, 32\}$. We empirically selected the optimal value for α by a grid search in $0 < \alpha < 1$ with 0.1 increments. For each downstream task, the best value of α sets to 0.7. All experiments were performed with three training epochs and using an NVIDIA V100 GPU.

4 Results and Discussion

Table 2 illustrates the debiasing performance comparison between Gender-tuning and two previous studies and their performance on downstream task datasets after debiasing. We report SEAT absolute effect size (e-size) on sentence templates of Terms/Names under different gender-domains provided by (Caliskan et al., 2017). Also, we report the results of debiasing when Gender-tuning perturbs the input example randomly (select randomly 5% of the input tokens for perturbing, not just gender words). The best results of the individual SEAT test effect size (e-size), average absolute e-size (lower is better), and accuracy performance (higher is better) are shown in bold font.

[2]https://github.com/huggingface/transformers

Table 2: Debiasing results in BERT and RoBERTa using SST-2, CoLA, and QNLI datasets. First six rows measure binary SEAT effect size (e-size) for sentence-level tests from (Caliskan et al., 2017). The seventh row presents the absolute value of average effect size (Avg. Abs. e-size). SEAT scores closer to 0 represent lower bias. Also, the eighth row shows the accuracy performance after debiasing. We compared our proposed method using the BERT with two recent embedding debiasing methods, Sent-D (Liang et al., 2020) and FairF (Cheng et al., 2021). The origin model has been implemented from huggingface. The Gender-tuning$_{random}$ masks the input example randomly, the Gender-tuning masks the gender-related words. Gender-tuning gains the lowest average bias in both models.

SST-2	**BERT**					**RoBERTa**		
	Origin	Sent-D	FairF	Gender-tuning_random	Gender-tuning	Origin	Gender-tuning_random	Gender-tuning
Names, Career/Family	0.28	0.10	0.21	0.46	**0.03**	**0.07**	0.08	0.14
Terms, Career/Family	0.18	0.05	0.37	**0.03**	0.16	0.33	0.44	**0.01**
Terms, Math/Art	0.49	0.22	0.26	**0.05**	0.39	1.32	1.25	**0.57**
Names, Math/Art	0.59	0.75	**0.09**	0.65	0.31	1.34	1.12	**1.11**
Terms, Science/Art	0.14	0.08	0.12	0.42	**0.08**	0.25	**0.12**	0.47
Names, Science/Art	**0.02**	0.04	0.05	0.38	0.10	**0.47**	0.62	**0.47**
Avg. Abs. e-size	0.283	0.212	0.182	0.331	**0.178**	0.630	0.605	**0.461**
Accuracy	91.97	89.10	91.60	**92.66**	92.10	93.57	**93.92**	93.69
CoLA								
Names, Career/Family	0.45	0.14	**0.03**	0.34	0.10	0.29	0.15	**0.05**
Terms, Career/Family	0.08	0.18	0.11	0.15	**0.03**	0.26	0.08	**0.00**
Terms, Math/Art	0.73	0.31	**0.09**	0.55	0.53	0.06	**0.02**	0.15
Names, Math/Art	0.97	0.30	**0.10**	0.72	0.24	**0.06**	0.25	0.07
Terms, Science/Art	0.41	0.16	0.24	**0.05**	0.37	**0.32**	0.57	0.70
Names, Science/Art	0.33	0.19	0.12	0.28	**0.07**	0.27	0.14	**0.03**
Avg. Abs. e-size	0.495	.217	**0.120**	0.343	0.223	0.210	0.201	**0.166**
Accuracy	56.51	55.40	56.50	**56.85**	56.60	57.35	57.55	**58.54**
QNLI								
Names, Career/Family	0.11	0.05	0.10	**0.01**	0.02	**0.04**	0.38	0.17
Terms, Career/Family	0.35	**0.004**	0.20	0.13	0.04	0.22	0.10	**0.04**
Terms, Math/Art	0.09	**0.08**	0.32	0.30	**0.08**	0.53	0.16	**0.09**
Names, Math/Art	0.28	0.62	0.28	0.23	**0.16**	0.48	0.06	**0.03**
Terms, Science/Art	0.34	0.71	0.24	0.25	**0.21**	**0.47**	0.57	0.53
Names, Science/Art	0.10	0.44	0.16	0.15	**0.04**	**0.36**	0.47	0.52
Avg. Abs. e-size	0.211	0.321	0.222	0.178	**0.091**	0.350	0.290	**0.230**
Accuracy	91.30	90.60	90.80	**91.61**	91.32	92.03	**92.51**	92.09

Compared with the original BERT and RoBERTa fine-tuning results (Table 2), Gender-tuning effectively reduces the average absolute effect size for both language models on all downstream tasks. However, compared with the previous debiasing methods on the BERT language model, Gender-tuning gains the smallest average effect size on SST-2 and QNLI. On the CoLA dataset, Gender-tuning on the BERT language model got the smallest SEAT effect size on the 'Terms, Career/Family' and 'Name, Science/Art' domains.

Moreover, in contrast with the previous debiasing methods that mostly hurt the language model performance on the downstream tasks, Gender-tuning improves the performance on downstream tasks. This means that the proposed method preserves the useful semantic information of the training data after debiasing. According to the Table 2, BERT model's results, Gender-tuning$_{random}$ obtains the best accuracy performance. This is because the size of the training examples in random perturbation is larger than when perturbing only the examples that contain the gender word(s). However, in the RoBERTa model, which is pre-trained on a significantly larger-scale of pre-training corpora than BERT, Gender-tuning notably improves performance accuracy.

4.1 Ablation

We conducted the ablation experiment to demonstrate the importance of computing the joint-loss for training Gender-tuning. For this aim, we perform Gender-tuning without using the joint-loss (**Gender-tuning**$_{no-Joint}$). In this case, only fine-tuning loss trains the Gender-tuning training processes (i.e., gender-

Table 3: Comparing the results from Gender-tuning$_{no-Joint}$ that uses only the fine-tuning loss with Origin models (from huggingface) and our proposed methodt (Gender-tuning). The results show that Gender-tuning achieved the least bias average value on all downstream task datasets in both models, BERT and RoBERTa.

SST-2	**BERT**			**RoBERTa**		
	Origin	Gender-tuning_no- Joint	Gender-tuning	Origin	Gender-tuning_no- Joint	Gender-tuning
Names, Career/Family	0.28	0.16	**0.03**	**0.07**	0.62	0.14
Terms, Career/Family	0.18	0.37	**0.16**	0.33	0.41	**0.01**
Terms, Math/Art	0.49	0.49	**0.39**	1.32	1.02	**0.57**
Names, Math/Art	0.59	0.56	**0.31**	1.34	**0.97**	1.11
Terms, Science/Art	0.14	0.32	**0.08**	0.25	**0.00**	0.47
Names, Science/Art	**0.02**	0.47	0.10	0.47	0.56	**0.46**
Avg. Abs. e-size	0.283	0.395	**0.178**	0.630	0.596	**0.461**
Accuracy	91.97	**92.66**	92.10	93.57	92.54	**93.69**
CoLA						
Names, Career/Family	0.45	**0.04**	0.1	0.29	0.16	**0.05**
Terms, Career/Family	0.08	0.11	**0.03**	0.26	0.11	**0.00**
Terms, Math/Art	0.73	0.96	**0.53**	**0.06**	0.29	0.15
Names, Math/Art	0.97	0.82	**0.24**	**0.06**	0.87	**0.07**
Terms, Science/Art	0.41	**0.19**	0.37	**0.32**	0.80	0.70
Names, Science/Art	0.33	0.32	**0.07**	0.27	0.88	**0.03**
Avg. Abs. e-size	0.495	0.406	**0.223**	0.210	0.518	**0.166**
Accuracy	56.51	**56.70**	56.60	57.35	57.27	**58.54**
QNLI						
Names, Career/Family	0.11	0.15	**0.02**	**0.04**	0.14	0.17
Terms, Career/Family	0.35	0.41	**0.04**	0.22	0.11	**0.04**
Terms, Math/Art	0.09	**0.03**	0.08	0.53	0.62	**0.09**
Names, Math/Art	0.28	**0.04**	0.16	0.48	0.42	**0.03**
Terms, Science/Art	0.34	0.27	**0.21**	**0.47**	0.50	0.53
Names, Science/Art	0.10	0.11	**0.04**	0.36	**0.20**	0.52
Avg. Abs. e-size	0.211	0.168	**0.091**	0.350	0.331	**0.230**
Accuracy	91.30	91.28	**91.32**	92.03	91.69	**92.09**

word perturbation and fine-tuning). In Table 3, we report the gender biases scores for **Gender-tuning**$_{no-Joint}$ and compare it with (1) **Origin**; original fine-tuning model, and (2) **Gender-tuning**; The proposed approach. The results prove the importance of the joint-loss as we mentioned earlier in Section 3.

In both BERT and RoBERTa models, results illustrate that Gender-tuning is more effective for reducing the average gender bias than Gender-tuning$_{no-Joint}$ using only fine-loss. Also, in most of the gender domains, Gender-tuning gains the smallest SEAT absolute effect size compared to the original model and Gender-tuning$_{no-Joint}$, especially in the BERT model. Indeed, in RoBERTa, Gender-tuning improves the pre-trained model performance noticeably. In the BERT model, Gender-tuning$_{no-Joint}$ achieves marginally higher performance accuracy on the downstream tasks except the CoLA dataset. This success achieves by cooperation between two training steps' error in Gender-tuning and computing joint-loss. Gender-tuning$_{no-Joint}$ does not update the masked and pre-trained language models when the output of the fine-tuning classification is correct ($\mathcal{L}_{fine} \approx 0$). Even though the correct output likely bases on the gender biases existing in the masked or pre-trained language model.

Moreover, through the comprehensive experiments, we observed that sometimes example perturbation changes the input example in a way that is not conceptually related to the ground-truth label. For example, when input example is :{[He] is a wonderful [actor]" (label: positive)}, one of the possibility for perturbation result can be {"[She] is a wonderful [nightmare]" (label: positive)}. Based on this example, let's assume that the Gender-tuning classification output becomes correct, and consequently, the fine-loss is close to zero. Thus the language models are not updated and learn the wrong concept (the sentiment of the perturbed example changes to negative after perturbation, but Gender-tuning assigns it to the ground-truth label, which is positive.). In this case, joint-loss survives the language model by considering the perturbation loss for

training iteration. When the predicted word is far from gender-related, the perturbation error is big enough to retrain the language model.

5 Related Works

Social biases have recently been recognized as a critical issue in pre-trained language models. Some studies proposed different solutions for mitigating social biases in pre-trained language models, e.g., (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018; May et al., 2019; Kurita et al., 2019; Sheng et al., 2019; Basta et al., 2019; Webster et al., 2020; Gehman et al., 2020; Abid et al., 2021). These solutions can be categorized into two groups (Blodgett et al., 2020): debiasing database and debiasing embedding.

Debiasing Database; The most straightforward approach for reducing the social biases in the training corpora is dataset(s) bias-neutralization. In this way, the training corpus is directly re-balanced by swapping or removing bias-related words and counterfactual data augmentation (CDA) (Zmigrod et al., 2019; Dinan et al., 2020; Webster et al., 2020; Dev et al., 2020; Barikeri et al., 2021). Also, (Gehman et al., 2020) proposed domain-adaptive pre-training on unbiased corpora. Although the results showed these proposed methods mitigated the social biases in the pre-trained models, they need to perform the retraining on a larger scale of the corpora. For example, webster2020measuring proposed a CDA that needs an additional 100k steps of training on the augmented dataset. Data augmentation and collecting a large-scale unbiased corpus are both computationally costly.

Debiasing Embedding; There are several solutions for debiasing static word embedding (Bolukbasi et al., 2016; Kaneko & Bollegala, 2019; Manzini et al., 2019; Ravfogel et al., 2020) and debiasing contextualized word-embedding (Caliskan et al., 2017; Brunet et al., 2019) and sentence-embedding (Liang et al., 2020; Cheng et al., 2021). Compared to debiasing static word embedding, where the semantic representation of a word is limited to a single vector, contextualized word/sentence embedding models are more challenging (Kaneko & Bollegala, 2019). Since the key to the pre-trained language models' success is due to powerful embedding layers (Liang et al., 2020), debiasing embedding might affect transferring of the accurate information and performance of these models on the downstream tasks. Also, they need some pre-training for debiasing the embedding layer before fine-tuning on downstream tasks.

In this study, we developed the fine-tuning process by adding a gender word perturbation and using a joint-loss for training to avoid social biases propagation when training on the downstream tasks and reduce the biases scores in pre-trained language models. Thus our proposed approach is applicable to debiasing any pre-trained language models that work with the original fine-tuning. Indeed, Gender-tuning solely uses the downstream task dataset for debiasing the pre-trained language models. The results of a comprehensive experiment show that Gender-tuning effectively reduces the gender biases scores in pre-trained language models.

6 Conclusion

We proposed a novel debiasing approach for pre-trained language models by empowering the fine-tuning. In this study, we evaluated our proposed method on gender biases and named it *Gender-tuning*. Gender-tuning aggregates gender-word perturbation and fine-tuning for debiasing the pre-trained model on the downstream task dataset. Then an aggregation loss from these two steps is used for training iterations. The comprehensive experiments prove that Gender-tuning effectively reduces gender-bias scores while preserving semantic information in the pre-trained language models. Thus Gender-tuning improves the performance on downstream tasks as well. The key advantage of our approach is using the fine-tuning setting that allows the learning process to be carried out without the need for additional training data and the pre-training process. Also, it makes Gender-tuning as a plug-and-play debiasing tool for any pre-trained language models.

REFERENCES

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. *arXiv preprint arXiv:2101.05783*, 2021.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*, 2021.

Christine Basta, Marta R Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 33–39, 2019.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *NAACL HLT 2019*, pp. 7, 2019.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pp. 803–811. PMLR, 2019.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*, 2021.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. 2020.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7659–7666, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8173–8188, 2020.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 3356–3369, 2020.

Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1641–1650, 2019.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 166–172, 2019.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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. 615–621, 2019.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *NAACL-HLT (1)*, 2019.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2016.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *arXiv preprint arXiv:2103.00453*, 2021.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. 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. 3407–3412, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *EMNLP*, 2018.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661, 2019.