Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models

Somayeh Ghanbarzadeh
University of North Texas
`somayehghanbarzadeh@my.unt.edu`

Yan Huang
University of North Texas
`yan.huang@unt.edu`

Hamid Palangi[1] **Radames Cruz Moreno**[2] **Hamed Khanpour** [3]
Microsoft Research   Microsoft Research   Microsoft Research
`{hpalangi`[1]`,radames.cruz`[2]`,hamed.khanpour`[3]`}@microsoft.com`

Abstract

Recent studies have revealed that the widely-used Pre-trained Language Models (PLMs) propagate societal biases from the large unmoderated pre-training corpora. Existing solutions require debiasing training processes and datasets for debiasing, which are resource-intensive and costly. Furthermore, these methods hurt the PLMs' performance on downstream tasks. In this study, we propose *Gender-tuning*, which debiases the PLMs through fine-tuning on downstream tasks' datasets. For this aim, Gender-tuning integrates Masked Language Modeling (MLM) training objectives into fine-tuning's training process. Comprehensive experiments show that Gender-tuning outperforms the state-of-the-art baselines in terms of average gender bias scores in PLMs while improving PLMs' performance on downstream tasks solely using the downstream tasks' dataset. Also, Gender-tuning is a deployable debiasing tool for any PLM that works with original fine-tuning.

1 Introduction

Pre-trained Language Models (PLMs) have achieved state-of-the-art performance across various 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 on large-scale corpora, which is collected from unmoderated sources such as the internet. Prior studies (Caliskan et al., 2017; Zhao et al., 2018; May et al., 2019; Kurita et al., 2019; Gehman et al., 2020) have shown that PLMs capture a significant amount of social biases existing in the pre-training corpus. For instance, they showed that the PLMs learned that the word "he" is closer to the word "engineer" because of the frequent co-occurrence of this combination in the training corpora, which is known as social gender biases. Since PLMs are increasingly deployed in real-world scenarios, there is a serious concern that they propagate discriminative prediction and unfairness.

Several solutions for mitigating the social biases have been proposed, including: using banned word lists (Raffel et al., 2020), building deliberated training datasets (Bender et al., 2021), balancing the biased and unbiased terms in the training dataset (Dixon et al., 2018; Bordia and 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, they tend to limit the PLMs' ability (Meade et al., 2022). For example, the banned words solution prevent gaining knowledge of topics related to banned words. Also, some of them hurt the PLMs' performance on downstream tasks. Furthermore, dataset curation and pre-training are two resource-intensive tasks needed for most of the above solutions (Schick et al., 2021).

In this study, we address the challenges mentioned above by proposing an effective approach named *Gender-tuning* for debiasing the PLMs through fine-tuning on downstream tasks' datasets. For this goal, Gender-tuning perturbs the training examples by first finding the gender-words in the training examples based on a given gender-word list. Then Gender-tuning replaces them with the new words to interrupt the association between the gender-words and other words in the training examples (Table 1). Finally, Gender-tuning classifies the examples with the replaced words according to the original training examples' ground-truth labels to compute a joint loss from perturbation and classification for training the Gender-tuning. The key advantage of our method is integrating the debiasing process into the fine-tuning that allows the debiasing and fine-tuning to perform simultaneously. Thus, Gender-tuning does not require separate pre-training or additional training data. Also, this integration makes Gender-tuning a plug-and-

1	**Original example:**
	"[**he**] is at 22 a powerful [**actor**]."
	Perturbed examples:
	epoch 1 ⇒ "[**girl**] is at 22 a powerful [**UNK**]."
	epoch 2 ⇒ "[**boy**] is at 22 a powerful [**actor**]."
	epoch 3 ⇒ "[**She**] is at 22 a powerful [**actress**]."
2	**Original example:**
	"[**she**] beautifully chaperon the [**girls**] in the kitchen."
	Perturbed examples:
	epoch 1 ⇒ "[**lady**] beautifully chaperon the [**women**] in the kitchen."
	epoch 2 ⇒ "[**girl**] beautifully chaperon the [**boys**] in the kitchen."
	epoch 3 ⇒ "[**he**] beautifully chaperon the [**men**] in the kitchen."

Table 1: Some perturbed examples generated by Gender-tuning through three training epochs.

play debiasing tool for any PLMs that works with original fine-tuning.

To evaluate the effectiveness of our proposed method, we conducted comprehensive experiments following two state-of-the-art debiasing baselines: SENT_DEBIAS (Sent-D) (Liang et al., 2020) and FairFil (FairF) (Cheng et al., 2021). The results show that Gender-tuning outperforms both baselines in terms of the average gender-bias scores in the BERT model while improving its performance on the downstream tasks. In addition, we reported the performance of Gender-tuning applied to the RoBERTa that shows considerable improvement. Finally, our ablation studies demonstrate that all components of Gender-tuning, including two training phases and joint loss, play an essential role in achieving success.

2 Methodology

We propose a novel debiasing approach, named Gender-tuning (Figure 1), that performs the debiasing process and fine-tuning simultaneously on the downstream tasks' dataset. For this aim, Gender-tuning integrates two training objectives: 1) Masked Language Modeling (MLM) training objective for gender-word perturbation and 2) Fine-tuning for classification. In each training batch, Gender-tuning works as follows:
Gender-tuning uses MLM to perturb training examples by masking the existing gender-word(s)[1]. The MLM training objective is to predict masked token(s) with a mean cross-entropy loss that we denote as perturbation-loss ($\mathcal{L}_{perturb}$). The training examples with predicted tokens, called *gender-perturbed examples* (Table 1), are fed into fine-

---

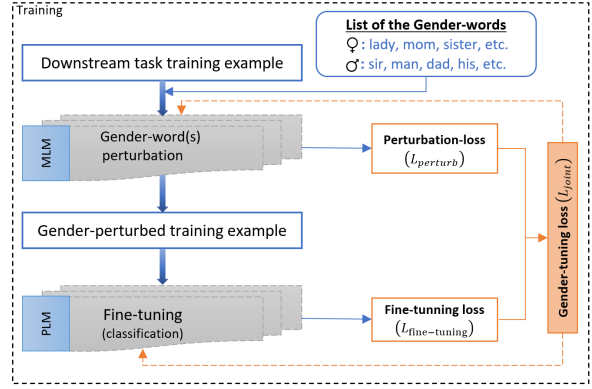[1]We use the feminine and masculine word lists created by (Zhao et al., 2018)

Figure 1: Illustration of Gender-tuning training process. MLM and PLM be trained based on Gender-tuning loss. The examples without any gender-word are directly fed to fine-tuning.

tuning to be classified according to the original examples' ground-truth label (y). Then $p_\theta(y' = y|\hat{x})$ is the fine-tuning classification function to predict the gender-perturbed example's label (y') based on the gender-perturbed example ($\hat{x}$) to compute the fine-tuning loss ($\mathcal{L}_{fine-tuning}$), where θ is the PLM's parameters for the fine-tuning. A weighted aggregation of the perturbation loss and fine-tuning loss, called joint-loss ($\mathcal{L}_{joint}$), is used for training the Gender-tuning as follows:

$$\mathcal{L}_{joint} = \alpha\,\mathcal{L}_{perturb} + (1-\alpha)\mathcal{L}_{fine-tuning} \quad (1)$$

where α is a weighting factor that is employed to adjust the contribution of the two training losses in computing the joint-loss.

The Gender-tuning training objective is to minimize joint-loss to ensure that the label of the perturbed example is the same as the label of the original training example. In the following, we present how joint-loss impacts the training process of Gender-tuning in each training batch:

Suppose the MLM predicts an incorrect token. For instance, the example: "the film affirms the power of the [actress]" changes to "the film affirms the power of the [trauma]". In this example, the predicted word [trauma] is a non-related gender-word that raises perturbation-loss value ($\mathcal{L}_{perturb}$ > 0). In this case, even if fine-tuning classifies the perturbed example correctly, joint-loss is still big enough to force Gender-tuning to continue training.

Also, suppose Gender-tuning creates social gender bias through gender perturbation. For instance, the example: "angry black [actor]" changes to "angry black [woman]" that "woman" and "actor" are not close semantically that raises perturbation-loss

SST-2	BERT					RoBERTa		
	Origin	Sent-D	FairF	Gender-tuning$_{random}$	Gender-tuning (ours)	Origin	Gender-tuning$_{random}$	Gender-tuning (ours)
Names, Career/Family	0.03	0.10	0.21	0.46	**0.03**	0.07	**0.08**	0.14
Terms, Career/Family	0.01	0.05	0.37	**0.03**	0.16	0.33	0.44	**0.01**
Terms, Math/Art	0.21	0.22	0.26	**0.05**	0.39	1.32	1.25	**0.57**
Names, Math/Art	1.15	0.75	**0.09**	0.65	0.31	1.34	1.12	**1.11**
Terms, Science/Art	0.10	0.08	0.12	0.42	**0.07**	0.25	**0.12**	0.47
Names, Science/Art	0.22	0.04	**0.005**	0.38	0.10	0.47	0.62	0.47
Avg. Abs. e-size	0.291	0.212	0.182	0.331	**0.176**	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	0.009	0.14	**0.03**	0.34	0.09	0.29	0.15	**0.05**
Terms, Career/Family	0.19	0.18	0.11	0.15	**0.03**	0.26	0.08	**0.00**
Terms, Math/Art	0.26	0.31	0.09	0.55	**0.08**	0.06	**0.02**	0.15
Names, Math/Art	0.15	0.30	**0.10**	0.72	0.24	0.06	0.25	**0.07**
Terms, Science/Art	0.42	0.16	0.24	**0.05**	0.07	0.32	**0.57**	0.70
Names, Science/Art	0.03	0.19	0.12	0.28	**0.07**	0.27	0.14	**0.03**
Avg. Abs. e-size	0.181	.217	0.120	0.343	**0.096**	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	0.26	0.05	0.10	**0.01**	0.02	0.04	0.38	**0.17**
Terms, Career/Family	0.15	**0.004**	0.20	0.13	0.04	0.22	0.10	**0.04**
Terms, Math/Art	0.58	**0.08**	0.32	0.30	**0.08**	0.53	0.16	**0.09**
Names, Math/Art	0.58	0.62	0.28	0.23	**0.16**	0.48	0.06	**0.03**
Terms, Science/Art	0.08	0.71	0.24	0.25	**0.21**	0.47	0.57	**0.53**
Names, Science/Art	0.52	0.44	0.16	0.15	**0.04**	0.36	0.47	0.52
Avg. Abs. e-size	0.365	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

Table 2: Comparing the debiasing performance of Gender-tuning and two state-of-the-art baselines. First six rows measure binary SEAT effect size (e-size; lower is better) for sentence-level tests from (Caliskan et al., 2017). The seventh row presents the average absolute e-size. The eighth row shows the classification accuracy on downstream tasks. The Gender-tuning$_{random}$ masks the input example randomly (not only gender-words). Gender-tuning gains the lowest average bias in both models and all datasets.

value ($\mathcal{L}_{perturb} > 0$). In this case, the output of the fine-tuning might be correct ($\mathcal{L}_{fine-tuning} \approx 0$) due to the PLMs' learned biases ("angry black woman" is a known gender/race bias). However, due to the big value of perturbation-loss, the join-loss is big enough to override fine-tuning results and forces Gender-tuning to continue training.

Moreover, we observed that sometimes example perturbation changes the concept/label of training examples. For instance, the input: "[He] is an excellent [actor] (label: positive)" changes to "[She] is a wonderful [murderer] (label: positive)", and fine-tuning classification output is correct ($\mathcal{L}_{fine-tuning} \approx 0$). In this example, the predicted word [murderer] is conceptually far from gender-related words [actor]. So, perturbation loss becomes significant, which creates a big value for joint-loss to force Gender-tuning to continue training. Finally, we found examples that MLM replaces the gender-word with the [UNK] token. In these examples, the perturbation-loss is close to zero ($\mathcal{L}_{perturb} \approx 0$) and the output of the fine-tuning classifier is incorrect ($\mathcal{L}_{fine-tuning} > 0$). In this case, the joint-loss is big enough to continue training and provide a new chance for MLM to predict a meaningful token instead of a [UNK]. More analysis of our perturbation strategy can be found in Section 4.1 and Table 3.

3 Experimental Setup

To evaluate our proposed method, we conduct experiments by following the evaluation process of the two state-of-the-art baselines (Sent-D and FairF) such as the bias evaluation metric (SEAT), applied PLMs, and downstream tasks' datasets.[2]

We report the SEAT effect size (e-size), average absolute e-size, and classification accuracy on downstream tasks for three different setups: 1) **Origin**: fine-tuning the PLMs on the downstream task datasets using huggingface transformers code (Wolf et al., 2020). 2) **Gender-tuning$_{random}$**: instead of replacing the gender-words in an training example, Gender-tuning$_{random}$ replaces a certain percentage of an input tokens randomly (5% of each input sequence). 3) **Gender-tuning**: the proposed method. We used the same hyperparameter for all three setups for a fair comparison.

4 Results and Discussion

Table 2 illustrates SEAT absolute effect size (e-size) (lower is better) on sentence templates of Terms/Names under different gender domains provided by (Caliskan et al., 2017), average absolute e-size (lower is better), and classification accuracy on downstream tasks (higher is better)

---

[2]Details of the baselines, bias evaluation metric, PLMs, datasets, and hyperparameters are presented in Appendix A.

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

Table 3: The illustration of the different types of perturbation outputs generated by Gender-tuning and their ground-truth label.

for three experiment setups (Section 3) and two state-of-the-art baselines. The results show that Gender-tuning outperforms the baselines regarding the average absolute effect size for both PLMs on all datasets. Also, in contrast with the baselines, Gender-tuning improves the accuracy of both PLMs on all downstream tasks. It shows that the proposed method preserves the useful semantic information of the training data after debiasing. The Gender-tuning$_{random}$ results show an inconsistent effect on the bias scores. Although Gender-tuning$_{random}$ improves the PLMs' accuracy on the downstream tasks, it significantly magnifies the bias score in the BERT model on SST-2 and CoLA. Also, it slightly reduces the average bias score in the RoBERTa on all datasets and in BERT on the QNLI.

4.1 Perturbation Analysis

The PLMs achieved state-of-the-art performance on the downstream tasks datasets by applying the MLM for the example perturbation in pre-training phase. Thus we hypothesize that the MLM can generate realistic gender-perturbed examples that can considerably modify the gender relation between the input tokens without affecting the label. However, there is a concern that the pre-trained MLM transfers the gender bias through the perturbation process.

To address this concern, we investigate the predicted tokens that the pre-trained MLM 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 3:

- **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 3, Gender-tuning does not make a meaningful change in identical and same-gender perturbation. These examples likely conform to the gender biases in the MLM. 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 MLM is to get the correct output from fine-tuning step and joint-loss close to zero. This issue stops the MLM and fine-tuning model from further update. However, joint-loss plays an essential role

SST-2	BERT				RoBERTa			
	Origin	Gender-tuning/ no-joint-train	Gender-tuning/ no-joint-loss	Gender-tuning (ours)	Origin	Gender-tuning/ no-joint-train	Gender-tuning/ no-joint-loss	Gender-tuning (ours)
Names, Career/Family	0.03	0.22	0.16	**0.03**	0.07	0.18	0.62	**0.14**
Terms, Career/Family	0.01	0.31	0.37	**0.16**	0.33	0.09	0.41	**0.01**
Terms, Math/Art	0.21	0.75	0.49	**0.39**	1.32	0.99	1.02	**0.57**
Names, Math/Art	1.15	0.55	0.56	**0.31**	1.34	**0.92**	0.97	1.11
Terms, Science/Art	0.10	0.01	0.32	**0.07**	0.25	0.76	**0.00**	0.47
Names, Science/Art	0.22	**0.07**	0.47	0.10	0.47	0.76	0.56	**0.47**
Avg. Abs. e-size	0.291	0.318	0.395	**0.176**	0.630	0.616	0.596	**0.461**
Accuracy	91.97	**92.88**	92.66	92.10	93.57	**94.38**	92.54	93.69
CoLA								
Names, Career/Family	0.09	0.37	**0.04**	0.09	0.29	0.07	0.16	**0.05**
Terms, Career/Family	0.19	0.06	0.11	**0.03**	0.26	0.16	0.11	**0.00**
Terms, Math/Art	0.26	0.89	0.96	**0.08**	0.06	0.41	0.29	**0.15**
Names, Math/Art	0.15	1.03	0.82	**0.24**	0.06	0.22	0.87	**0.07**
Terms, Science/Art	0.42	0.47	0.19	**0.07**	0.32	**0.42**	0.80	0.70
Names, Science/Art	0.03	0.49	0.32	**0.07**	0.27	0.36	0.88	**0.03**
Avg. Abs. e-size	0.181	0.551	0.406	**0.096**	0.210	0.273	0.518	**0.166**
Accuracy	56.51	56.32	**56.70**	56.60	57.35	**62.11**	57.27	58.54
QNLI								
Names, Career/Family	0.26	0.03	0.15	**0.02**	0.04	**0.12**	0.14	0.17
Terms, Career/Family	0.15	0.20	0.41	**0.04**	0.22	0.31	0.11	**0.04**
Terms, Math/Art	0.58	0.47	**0.03**	0.08	0.53	0.50	0.62	**0.09**
Names, Math/Art	0.58	0.94	**0.04**	0.16	0.48	0.38	0.42	**0.03**
Terms, Science/Art	0.08	**0.12**	0.27	0.21	0.47	**0.25**	0.50	0.53
Names, Science/Art	0.52	0.54	0.11	**0.04**	0.36	**0.03**	0.20	0.52
Avg. Abs. e-size	0.365	0.383	0.168	**0.091**	0.350	0.265	0.331	**0.230**
Accuracy	91.30	**91.57**	91.28	91.32	92.03	**92.58**	91.69	92.09

Table 4: Comparing the debiasing performance of two ablation experiments and Gender-tuning (ours) on three downstream task datasets. The results show that Gender-tuning achieved the least average bias score and consistently improved the classification accuracy.

in alleviating learning gender bias from identical and same-gender perturbations.

To clarify the role of joint-loss in overcoming above 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 large enough to update the language models 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 MLM as well as the PLM.

5 Ablation

We conduct the ablation experiments to demonstrate the effectiveness of Gender-tuning components, including 1) joint-training process and 2) joint-loss in Gender-tuning's debiasing performance (Table 4). The experiments are as follows: 1) **Gender-tuning**$_{no-joint-training}$: first we used MLM to train the PLM through the gender-word perturbation on downstream task datasets. Then we fine-tuned the PLM on the downstream task dataset. 2) **Gender-tuning**$_{no-joint-loss}$: we train Gender-tuning based on only fine-tuning loss.

In both PLMs, results illustrate that Gender-tuning is more effective for reducing the average gender bias than in two ablation experiments. The two ablation experiments magnify the bias scores noticeably, while Gender-tuning gains the smallest SEAT absolute effect size, especially in the BERT model. Results also show that the ablation experiment setups that do not benefit from joint-loss cannot update the MLM and PLM when the output of the fine-tuning classification is correct ($\mathcal{L}_{fine-tuning} \approx 0$), even though the correct output likely bases on the gender biases in the PLMs.

6 Conclusion

We propose a novel approach for debiasing PLMs through fine-tuning on downstream tasks' datasets. The proposed method is an aggregation of bias-word perturbation using MLM and fine-tuning classification. In this study, we evaluated our proposed method on gender biases and named it *Gender-tuning*. Comprehensive experiments prove that Gender-tuning outperforms two state-of-the-art debiasing methods while improving the performance of the PLMs on downstream tasks. The key advantage of our approach is using the fine-tuning setting that allows the training process to be carried out without needing additional training processes or datasets. Also, it makes Gender-tuning a plug-and-play debiasing tool deployable to any PLMs.

7 Limitation

Although Gender-tuning succeeds in reducing the gender bias scores in the pre-trained language models, there are some limitations to performing debiasing. Gender-tuning only works on gender-related words list. Thus Gender-tuning cannot cover the probable gender biases that do not exist in its' list. We defer the gender-related word list modification to future research. All our experiments ran on English language texts with English gender-word morphology.

References

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

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

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

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

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

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

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

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

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

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

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

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

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

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. 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)*, pages 615–621.

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

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898.

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

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

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

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

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*, pages 353–355.

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

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. 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*, pages 1651–1661.

A Appendix

A.1 Baselines

For comparison purposes, we chose two state-of-the-art baselines which focus on debiasing sentence-level pre-trained text encoders in PLMs.

A.1.1 SENT-DEBIAS

SENT-DEBIAS (Liang et al., 2020) is an extension of the HARD-DEBIAS method (Bolukbasi et al., 2016) to debias sentences for both binary and multi-class bias attributes spanning gender and religion. The key advantage of Sent-D is the contextualization step in which bias-attribute words are converted into bias-attribute sentences by using a diverse set of sentence templates from text corpora. Sent-D is a four-step process that involves: identifying words that exhibit biased attributes, contextualizing them in sentences that contain these biases, creating sentence representations, estimating the subspace of the bias represented in the sentences, and debiasing general sentences by removing the projection onto this subspace.

A.1.2 FairFil

FairF (Cheng et al., 2021) is the first neural debiasing method for pretrained sentence encoders. For a given pretrained encoder, FairF learns a fair filter (FairFil) network, whose inputs are the original embedding of the encoder, and outputs are the debiased embedding. Inspired by the multi-view contrastive learning (Chen et al., 2020), for each training sentence, FairF first generates an augmentation that has the same semantic meaning but in a different potential bias direction. FairFil is contrastively trained by maximizing the mutual information between the debiased embeddings of the original sentences and corresponding augmentations. To further eliminate bias from sensitive words in sentences, FairF uses debiasing regularizer, which minimizes the mutual information between debiased embeddings and the sensitive words' embeddings.

A.2 Bias Evaluation Metric

Following the prior studies (Sent-D and FairF), 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 $female$ 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:

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

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

A.3 PLMs

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.

A.4 Datasets

We conducted empirical studies on the following three tasks from the GLUE[3] 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]https://gluebenchmark.com/tasks

A.5 Hyperparameters

The hyperparameters of the models, except batch size, are set to their default[4] 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.

A.6 Related Works

Debiasing Database; The most straightforward approach for reducing the social biases in the training corpora is 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 be re-trained on a larger scale of the corpora. For example, Webster et al. (2020) 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 and 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 and 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.

---

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

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.