IMPROVING GENDER FAIRNESS OF PRE-TRAINED LANGUAGE MODELS WITHOUT CATASTROPHIC FORGETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although pre-trained language models, such as BERT, achieve state-of-art performance in many language understanding tasks, they have been demonstrated to inherit strong gender bias from its training data. Existing studies addressing the gender bias issue of pre-trained models, usually recollect and build gender-neutral data on their own and conduct a second phase pre-training on the released pre-trained model with such data. However, given the limited size of the gender-neutral data and its potential distributional mismatch with the original pre-training data, catastrophic forgetting would occur during the second-phase pre-training. Forgetting on the original training data may damage the model's downstream performance to a large margin. In this work, we first empirically show that even if the gender-neutral data for second-phase pre-training comes from the original training data, catastrophic forgetting still occurs if the size of gender-neutral data is smaller than that of original training data. Then, we propose a new method, GEnder Equality Prompt (GEEP), to improve gender fairness of pre-trained models without forgetting. GEEP learns gender-related prompts to reduce gender bias, conditioned on frozen language models. Since all pre-trained parameters are frozen, forgetting on information from the original training data can be alleviated to the most extent. Then GEEP trains new embeddings of profession names as gender equality prompts conditioned on the frozen model. This makes GEEP more effective at debiasing as well. Because gender bias from previous data embedded in profession embeddings is already removed when they are re-intialized in GEEP before second-phase pre-training starts. Empirical results show that GEEP not only achieves state-of-the-art performances on gender debiasing in various applications such as pronoun predicting and coreference resolution, but also achieves comparable results on general downstream tasks such as GLUE with original pre-trained models without much forgetting.

1 INTRODUCTION

Pre-trained language models, e.g., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have shown competitive performance in a wide variety of NLP downstream applications. However, such models are often prone to exhibit gender bias (de Vassimon Manela et al., 2021; Zhao et al., 2019; Webster et al., 2020), due to its large scale unsupervised training data from the web (Liu et al., 2019; Brown et al., 2020). Gender bias refers to unbalanced model behaviors with respect to a specific gender (Cheng et al., 2020). Naturally, a model inherits gender bias from a biased training corpus. For example, studies show that BookCorpus and English Wikipedia data, which are commonly used to train the BERT model, suffer from gender imbalance (Tan & Celis, 2019; Wagner et al., 2016). Although significant advances have been made in alleviating gender bias in traditional NLP fields such as standard word embeddings(Bolukbasi et al., 2016; Zhao et al., 2018a;b), there are limitations in addressing this issue for pre-trained language models. Given the large amount, diversity, and opacity of the pre-training data, even if we have access to a pre-trained language model that is released to the general public such as GPT-3, getting access to the original pre-training data and manually filtering out all the bias-related text seems impossible.

Due to this limitation, existing studies trying to address the gender bias issue of pre-trained models, usually recollect and build gender-neutral data on their own and conduct a second phase pre-training on the released pre-trained model with such data (Webster et al., 2020; de Vassimon Manela et al., 2021). However, given the limited size of the gender-neutral data and its potential distributional mismatch with the original pre-training data, *catastrophic forgetting* problem can occur during the second-phase pre-training of such methods. Catastrophic forgetting (Kirkpatrick et al., 2017) is a long-standing problem in deep learning, which illustrates the tendency of a neural network to forget previously learned information upon learning new information. When it comes to second-phase pre-training of the pre-trained model, using the limited gender-neutral data to keep updating the entire massive model could potentially make the model forget the diverse information from the original pre-training data. Since the diversity and amount of training data are closely relevant to the pre-trained model's performance on downstream tasks(Liu et al., 2019), forgetting on the original training data may damage the model's downstream performance to a large margin.

In this paper, we first empirically show that even if the gender-neutral data for second-phase pre-training comes from the original training data set, the catastrophic forgetting problem still occurs if the size of debiased data is smaller than that of original training data. To build the gender-neutral data set for second-phase pre-training, we firstly filter English Wikipedia text to get sentences with occupations and professions, such as "nurse". Then, for each of these sentences, we anonymize person entities and swap the gender-related terms in it, such as "he" to "she", to form new sentences. Finally, we mix these new sentences together with the original occupation-related sentences as the gender-neutral data for second-phase pre-training. The size of the gender-neutral data is 78.3% of the Wikipedia and Book Corpus data, the original pre-training data of the BERT base model. We find that although the two data sets couldn't be more similar to each other and that the gender-neutral data for second-phase pre-training is not significantly smaller than the original data, the model's performance on downstream tasks such as GLUE (Wang et al., 2018), still drops with a considerable margin after second-phase pre-training.

Therefore, we propose a new method, GEnder Equality Prompt (GEEP), to alleviate gender bias of pre-trained models without forgetting. At second-phase pre-training with the gender-neutral data, GEEP updates gender-related prompts to reduce gender bias, conditioned on frozen pre-trained models. Specifically, inspired by recent prompt-tuning methods (Lester et al., 2021) for fine-tuning large pre-trained models, GEEP freezes all original parameters of the pre-trained model and only updates the newly extended parameters as gender equality prompts. Since all the pre-trained parameters are frozen, the forgetting of information from the original training data can be alleviated to the most extent. As for the gender equality prompts, different from prompt-tuning methods which add new trainable parameters as prompts without defining what they are, GEEP trains new word/token embeddings of profession names as gender equality prompts. Since gender bias issue is most prominent on profession names, training new embeddings for them makes GEEP more effective at debiasing than other second-phase pre-training methods as well. Since the embeddings of profession names are newly re-initialized when debiasing training starts, gender bias from previous data that is embedded in such representations is already removed before second-phase pre-training. Therefore, GEEP doesn't have to train the model to find and fix bias from scratch, which makes the debiasing faster. Empirical results show that GEEP not only achieves state-of-the-art performances on gender debiasing in various applications such as pronoun predicting and coreference resolution, but also achieves comparable results on general downstream tasks such as GLUE with original pre-trained models.

2 Related Work

In this section, we review relevant work on gender bias identification and mitigation for pre-trained word embeddings and language models.

Standard word embedding models, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), provide geometrical encodings of words from their co-occurrence in different documents. Despite the popularity of these methods, studies show that such word embeddings are often prone to exhibit gender bias (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018b; Gonen & Goldberg, 2019; Sun et al., 2019; Garg et al., 2018; Zhao et al., 2018a). To mitigate gender bias, Bolukbasi et al. (2016) leverage a group of gender-specific words such as "she" and "he', to define a gender subspace and neutralize embeddings of gender neural occupation words in this subspace.

Zhao et al. (2018b) propose Gender-Neutral Global Vectors (GN-GloVe) learning scheme to keep protected attributes in a certain dimension and neutralize other dimensions of the word embedding vector. Although such gender sub-space methods show effectiveness at debiasing pre-trained word embeddings, they can not be directly applied to pre-trained language models. Because the context representation space of entire pre-trained models are more dynamic and a specific subspace for fairness cannot be easily defined.

Recent work on gender fairness of pre-trained language models, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), mostly focus on showing and measuring the gender bias embedded in such models(Zhao et al., 2019; Tan & Celis, 2019). These studies propose metrics to quantify gender bias in pre-trained language models (de Vassimon Manela et al., 2021; Tan & Celis, 2019; Webster et al., 2018). Moreover, Kurita et al. (2019) define several template sentences and mask the professions (e.g. programmer) and gender-specific tokens (e.g. he and she) sequentially, and then measure the association between gender-specific tokens and attributes in the BERT model. In our work, we employ such methods to evaluate GEEP and baseline methods on improving gender fairness. Existing works focusing on mitigating gender bias of pre-trained models, usually collect and build gender-neutral data on their own and conduct a second phase pre-training on the released pre-trained model (Webster et al., 2020; de Vassimon Manela et al., 2021; Cheng et al., 2020). For example, Cheng et al. (2020) take advantage of such data augmentation methods and train a fair filter (FairFil) network to maximize the mutual information between the representations of the original sentences and their corresponding augmentations. However, to the best of our knowledge, none of such methods analyze their debiased models' downstream performance on general NLP tasks such as average GLUE and the potential forgetting issue of such sencond-phase pre-training methods. In this work, we are the first to demonstrate empirically that even if the gender-neutral data for second-phase pre-training comes from the original training data set, the debiased model's performance on general downstream tasks such as GLUE, still drops with a considerable margin after the second-phase pre-training. Then, given this phenomenon, we propose GEEP to alleviate gender bias of pre-trained models without forgetting.

3 Gender Bias in Pre-trained Language Models

In this section, we first describe architectural and training details of the current widely-used pre-trained language model, BERT, as preliminaries of our method. Then, we identify how severe gender bias issue is in public available pre-trained BERT.

3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a multi-layer bidirectional transformer encoder that maps a sequence of token embeddings and positional embeddings to the contextual representations of tokens (Devlin et al. (2019)). BERT is a stack of multiple transformer layers Vaswani et al. (2017). Each layer contains two sub-layers: 1) self-attention layer, and 2) position-wise fully connected feed-forward network, each followed by a residual connection and a layer normalization step. Self-attention, which is also referred to as "Scaled Dot-Product Attention", produces its output by calculating the scaled dot products of queries and keys as the coefficients of the values,

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V. \tag{1}$$

Q (Query), K (Key), V (Value) are the hidden representations outputted from the previous layer and d is the dimension of the hidden representations To give the attention layer multiple representation subspaces and expand the model's ability to focus on different positions, the self-attention layer of transformers is extended to a multi-headed attention mechanism:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_H)W^O \tag{2}$$

$$head_k = Attention(QW_k^Q, KW_k^K, VW_k^V) \tag{3}$$

where $W_k^Q \in \mathbb{R}^{d\times d_K}, W_k^K \in \mathbb{R}^{d\times d_K}, W_k^V \in \mathbb{R}^{d\times d_V}$ are projection matrices. H is the number of heads and d_K and d_V are the dimensions of the key and value, respectively. Th outputs of the multi-headed attention layer are fed to a fully connected feed-forward network (FFN). The FFN usually

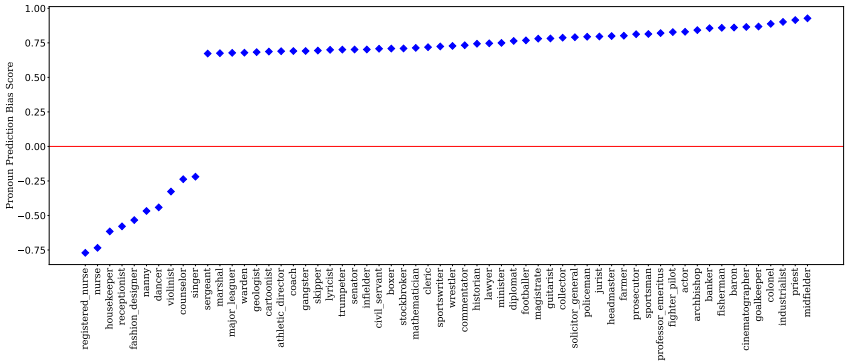


Figure 1: An example of gender bias in 60 most biased profession words in BERT-base model. For each profession, we measure the difference between the probability of filling the masked pronoun in each template sentence with "he" and "she" tokens. Some words such as nurse (-0.73) and receptionist (-0.57) are supposed to be gender neutral by definition but BERT-base model consider them as female professions. On the other hand, lawyer (0.74) and prosecutor (0.81) are considered as jobs for male.

consists of two linear projections with a ReLU activation in between:

$$FFN(h_i) = \delta(h_1 W_1 + b_1)W_2 + b_2 \tag{4}$$

where W_1, W_2, b_1 and b_2 are parameters.

The released pre-trained BERT model is trained on the BooksCorpus (800M words) and English Wikipedia (2,500M words) corpus with two unsupervised objective functions: 1) *masked language modeling (MLM)* and 2) *next sentence prediction*. In masked language modeling, 15% of all tokens in each sequence are replaced with [MASK] token at random and the model attempts to predict the masked tokens based on the context of unmasked words in the sequence. The input of the BERT model is sequences of sentences. In the next sentence prediction task, the model learns to predict if the current sentence is subsequent of the previous sentence in the training corpus. For fine-tuning, the model is initialized with the pre-trained parameters, and a new classification head is added to the core model. Then, all of the parameters are fine-tuned using labeled data from the downstream tasks.

3.2 Identifying Gender Bias in Pre-trained Language Models

Different approaches have been proposed to quantify and analyze the gender bias in contextual language models (de Vassimon Manela et al., 2021; Webster et al., 2020; Kurita et al., 2019). For BERT, we choose one approach that can be directly applied to a model pre-trained with MLM without further fine-tuning. In this approach, we first define a template containing a pronoun and a profession. The profession is supposed to be gender-neutral while is currently viewed with gender bias to a large extent. By masking the pronoun, the model is queried to predict the pronouns at the masked position given the context, including the profession. Here is an example, [MASK] is a registered nurse. The difference between the probabilities of filling the masked position in each sentence with "he" and "she", is used to show gender bias in the model,

$$\text{Pronoun Prediction Bias Score} = \text{Prob}(\textbf{"he"}) - \text{Prob}(\textbf{"she"}). \tag{5}$$

To assess fairness in BERT model, we consider 303 of professions used by Bolukbasi et al. (2016). In our study, we analyze a public available pre-trained BERT-Base model [1] that contains 12 layers, 768 hidden nodes, 12 heads, and 110M parameters. Figure 1 shows gender bias of 60 of such professions in BERT-base model. Positive values mean that the professions are biased towards male and vice versa. As the plots show, the contextual representations of professions in BERT-base model exhibits strong gender bias. Professions such as nurse and housekeeper are viewed as jobs for females while surgeon and mathematicians are assumed male.

[1] https://github.com/google-research/bert

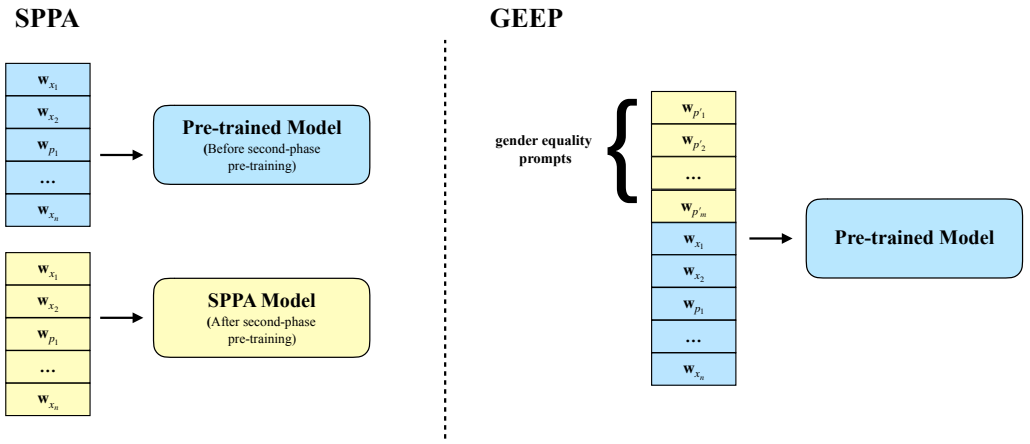


Figure 2: Difference between SPPA and GEEP methods. Blue boxes represent the parameters of the pre-trained model before any further training and yellow boxes show updated parameters during second-phase pre-training (SPPA). SPPA requires updating all the pre-trained model's parameters. GEEP method only requires initializing the profession words embeddings $\mathbf{w}_{p\prime_1}, \mathbf{w}_{p\prime_2}...., \mathbf{w}_{p\prime_m}$ randomly and updating them during second-phase pre-training while freezing all the pre-trained model's parameters.

4 Improving Gender Fairness without Forgetting

In this section, we describe in detail how the proposed method, GEnder Equality Prompt (GEEP) improves gender fairness of pre-trained models without forgetting.

4.1 Gender-neutral data collection

First, to construct data with proportionate numbers of references to male and female genders, we replicate the data augmentation method by Zhao et al. (2018a) on the English Wikipedia corpus which the BERT model is pre-trained on. We filter the dataset for sentences containing at least one profession that is supposed to be gender-neutral but generally viewed with gender bias, e.g., nurse, defined by Bolukbasi et al. (2016). We obtain $16,313,783$ sentences with such profession words. For each of these sentences, we swap the gendered terms with their opposite genders (such as "Man" $\rightarrow$"Woman", "he"$\rightarrow$"she", and vice-versa). Next, we use Named Entity Recognizer Lample et al. (2016) to identify person name entities in each sentence and replace them with anonymized entities, such as "ANON1". Our augmented dataset includes both the anonymized original and gender-swapped sentences, which is 78.3% of the original Wikipedia data.

After the gender-neutral data set is built, a common approach to mitigate gender bias in pre-trained language models is to conduct second-phase pre-training to update all model parameters with this dataset. We refer to such methods as *SPPA* (Second-Phase Pre-training for All parameters). As illustrated before, since all model parameters of the pre-trained model are updated in such methods with limited data, the various and diverse information that the model has captured from massive original training data might be forgotten. In Section 5, we empirically show that SPPA methods lead to forgetting issues even when the gender-neutral data for second-phase pre-training is collected from the original pre-training data.

4.2 Gender Equality Prompt Approach

To avoid catastrophic forgetting while mitigating gender bias in pre-trained language models, we propose GEnder Equality Prompt (GEEP). In GEEP, instead of updating all model parameters during second-phase pre-training, we freeze all previous model parameters in the pre-trained model and add additional trainable parameters. Since all the pre-trained parameters are frozen, the forgetting of information from the original training data can be alleviated to the most extent. Because gender bias issue is most prominent on profession names, GEEP adds new word/token embeddings of profes-

sion names as new trainable parameters. At second-phase pre-training, only the newly added token embeddings of profession names are updated with the gender-neutral data, conditioned on the original pre-trained model. We show the comparison between GEEP and other second-phase pre-training methods in Figure 2.

Let $\mathbf{X} = \{x_1, x_2, ...x_n\}$ denote the original vocabulary of the pre-trained model and $\mathbf{W}_x \in \mathbb{R}^{n \times d}$ be the token embedding matrix of the model with dimension of d. Given a set of m profession names, $\{p_1, p_2, ..., p_m\}$, we build an embedding matrix $\mathbf{W}_p \in \mathbb{R}^{m \times d}$ where the embedding of each token is initialized randomly. To obtain a integrated word embedding matrix, we concatenate $\mathbf{W}_x$ and W_p as $\mathbf{W}_{\text{emb}} = \text{Concat}(\mathbf{W}_x, \mathbf{W}_p)$. Then we use $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{(n+m) \times d}$ as the word embedding matrix for downstream tasks, as illustrated in Figure 2. Note that for both second-phase pre-training and fine-tuning, when the profession names are present in the input sequence, we only use and update their new embeddings in $\mathbf{W}_p$. Given the pre-trained model's frozen parameters $\mathbf{W}_{\text{base}}$, the objective function of second-phase pre-training of GEEP is,

$$\mathcal{L}(\mathbf{x}_{\text{masked}} | \mathbf{x}_{\text{context}}, \mathbf{W}_{\text{base}}) = \frac{1}{N_{\text{mask}}} (\sum_{t=1}^{N_{\text{mask}}} - \log p_\theta(x_t | \mathbf{x}_{\text{context}}, \mathbf{W}_{\text{base}})). \tag{6}$$

N_{mask} is the number of masked positions in the input sequence $\mathbf{x}$. With such an objective, $\mathbf{W}_p$ is updated with gender-neutral data. By training new embeddings $\mathbf{W}_p$ for biased professions, GEEP not only avoids forgetting, but also can be more effective at debiasing. Because in GEEP, the embeddings of profession names are newly re-initialized before debiasing training starts, so that gender bias from previous data embedded in such representations is already removed before second-phase pre-training. Therefore, GEEP doesn't have to train the model to find and fix bias from scratch, which can make the debiasing faster.

5 EXPERIMENTS

In this section, we first describe the experimental setup, including the model architecture, baselines, and hyper-parameters for second-phase pre-training. Then, we describe the downstream tasks and evaluation metrics that we use to test GEEP's performance on both gender fairness and alleviating forgetting. Finally, we present the results of GEEP and its baselines to show that GEEP achieves state-of-the-art performances on gender fairness tasks without hurting the pre-trained model's performance on general downstream tasks.

Table 1: The average accuracy of different models on Coreference Resolution task. The best results are in bold.

Data	BERT-base	BERT-SPPA	GEEP
Winogender	50	50.7	**62.9**
WSC	50.1	50.2	**50.5**
DPR/WSCR	50.7	50.9	**52.8**

5.1 EXPERIMENTAL SETUP

In our experiments, we mainly use publicly released BERT models as pre-trained models. A BERT-base model contains 12 transformer layers, 12 attention heads in each layer, and 110M parameters in total. Given a pre-trained BERT-base model, we compare GEEP with two main baselines. The first baseline is pre-trained BERT-base model without any further training. By comparing with this baseline, we can know to what extent GEEP overcomes the baseline's existing gender bias and whether GEEP hurts the baseline's general downstream performances by forgetting. The other important type of baselines is to conduct Second-Phase Pre-train to update All (SPPA) model parameters with the recollected gender-neutral data set. For a fair comparison with such methods, our SPPA baseline uses the exact same gender-neutral data set that we collect for GEEP (details in Section 3.2), and the same loss functions/hyper-parameters of BERT to further update all model parameters of the pre-trained BERT-base. For second-phase pre-training in GEEP and SPPA, we further train BERT-base for $10,000$ steps with our gender-neutral data. We use an AdamW optimizer with a learning rate of $2e-5$, max_seq_length of 128 and batch sizes $\in \{32, 256\}$. In GEEP method, we initialize the embedding of every profession prompt with a normal distribution and standard deviations of 0.2.

5.2 DOWNSTREAM TASKS AND EVALUATION METRICS

We conduct several experiments to show the effectiveness of GEEP in improving gender fairness and alleviating catastrophic forgetting compared to other methods. To assess gender fairness, we

conduct two main experiments: 1) pronoun prediction as described in section 3.2, and 2) coreference resolution. Coreference Resolution is the task of linking the pronouns with their references in a text. Studies show that BERT performance decreases in a text where the gender pronoun is female and the topic is biased towards the male gender (Kurita et al., 2019). To assess the performance of different models in pronoun coreference, we fine-tune our models with GAP dataset (Webster et al., 2018) and evaluate the performance of different models on three datasets:

- Winogender: This dataset includes $1,584$ sentences with three mentions: a profession, a participant, and a pronoun (where the pronoun is referred to either profession or pronoun).
- WSC: The Winograd Schema Challenge (WSC) incorporates 273 sentences used for commonsense reasoning for resolution (Levesque et al., 2012).
- DPR: The Definite Pronoun Resolution (DPR) corpus with 131 test sentences contains examples with two noun phrases and a pronoun or possessive adjective referring to one of the noun phrases (Rahman & Ng, 2012).

We fine-tune each model for one epoch with a train batch size of 64 and a learning rate of $5.0e-6$. To find the reference of each pronoun in the template sentences, we follow Kocijan et al. (2019) approach. Specifically, during the evaluation for every data set, in each sentence there are two candidate nouns (such as "nurse" or "surgeon") and a pronoun. The pronoun is replaced with a [MASK] token, and the model makes a prediction at the masked pronoun position from the two candidate nouns. In order to resolve a pronoun accurately, a model needs to overcome the biased link between gender and profession (e.g. a normative assumption that nurses are female) and instead make the decision based on the available linguistic cues. We report the prediction accuracy of all 3 methods on the aforementioned three data sets.

To evaluate how much each debiased model forgets after second-phase pre-training, we fine-tune the debiased models on all 8 GLUE tasks and report their performance on each GLUE task. The *General Language Understanding Evaluation (GLUE)* benchmark is a collection of eight tasks, widely used for evaluating the general language understanding capacity of pre-trained language models (Wang et al., 2018). For each task, we fine-tune our models using its train set and report the performance of the model on the development set. Our fine-tuning procedure follows the original BERT paper (Devlin et al., 2019). We consider a learning rate of $2e-5$ and batch size of 32. We fine-tune each model for 3 epochs, except in RTE task where we fine-tune for 10 epochs due to the small size of the dataset. Due to the variance in the performance of CoLA and RTE tasks, we report the average of the results for these tasks over five random initializations.

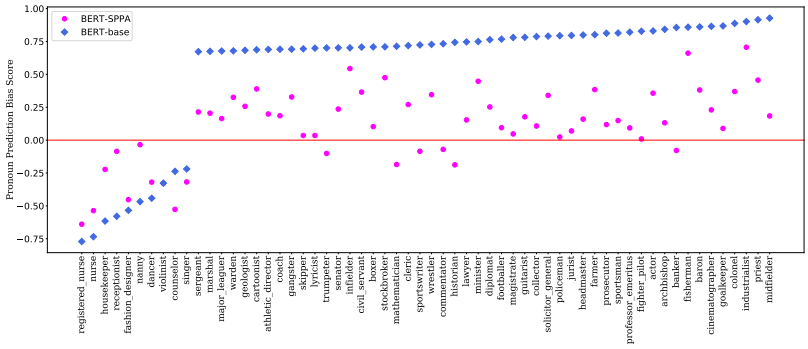
Table 2: GLUE results. For CoLA, we report accuracy/Matthews metrics. For STS-B, we report Pearson/Spearman. For all other tasks, we report accuracy score. The best results are in bold.

Task	BERT-base	BERT-SPPA	GEEP
CoLA	54.0/81.6	52.0/**81.2**	**53.0**/81.1
RTE	69.4	**69.8**	69.1
MRPC	85.7	84.1	**84.9**
STS-B	88.0/77.0	88.0/76.0	87.0/**77.0**
QQP	90	90	**90.4**
MNLI	84.3	84	**84.1**
QNLI	91.4	90	**91.3**
SST-2	93	92	**92.4**
AVG	83.0	82.3	**82.8**

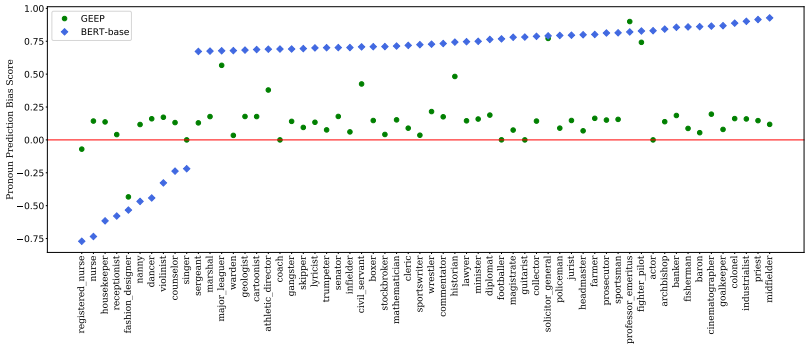
5.3 RESULTS

We first show the results for the pronoun prediction task. Figure 3 displays the pronoun prediction bias score (defined in Equation 5) of all methods for 60 biased professions defined in (Bolukbasi et al., 2016). Specifically, in both sub-figures, blue dots show the pronoun prediction bias score from BERT-base model for each profession. In Figure 3 (a), the pink dots are the bias scores from BERT-SPPA model. We can see from this sub-figure that compared with BERT-base, the bias scores from BERT-SPPA model are indeed closer to 0, indicating that BERT-SPPA can mitigate gender bias of such professions to some extent. In Figure 3 (b), the blue dots are the bias scores from GEEP model. Compared with both BERT-SPPA and BERT-base, GEEP's bias scores are significantly closer to 0, indicating that GEEP is more effective at removing gender bias from such baised professions compared with BERT-SPPA. Moreover, we also calculate the average absolute pronoun prediction

bias score for all 303 gender-neutral profession words in (Bolukbasi et al., 2016). We obtain 0.44 for BERT-base, 0.16 for BERT-SPPA and 0.13 for GEEP. GEEP model gets the lowest average bias with 70% reduction compared to the BERT-base model.



(a) Comparison between pronoun prediction bias in BERT-SPPA and BERT-base models



(b) Comparison between pronoun prediction bias in GEEP and BERT-base models

Figure 3: Difference between the probabilities of filling a masked pronoun with "he" and "she" tokens in the template sentences containing 60 most biased professions. GEEP method outperforms the two other methods. For example, the bias score for "nurse" token decreases from -0.7 in BERT-base to -0.5 in BERT-SPPA and 0.1 in GEEP model.

Then, we show the coreference resolution results of different models on three datasets in Table 1. Results show that GEEP model obtains the best accuracy compared to other models, specially in Wingender dataset where the candidate nouns are professions. We observe that the SPPA method also can help improve coreference resolution performance of the pre-trained model, but not as effective as GEEP.

Finally we show in Table 2 the performance of different models on 8 GLUE tasks, to see how severe the forgetting issue is in SPPA and GEEP. Compared with BERT, SPPA suffers from forgetting issue in the following 6 tasks out of the total 8 tasks, CoLA, MRPC, STS-B, MNLI, QNLI, and SST-2. As for the average GLUE score, SPPA is 0.7 point lower after its second-phase pre-training, which is not a small margin considering it is the average score of 8 tasks. GEEP mitigates the forgetting issue of SPPA in all sub-tasks except in RTE. GEEP also gets the average GLUE score of 82.8, which outperforms SPPA and is similar to the original GLUE score of the pre-trained BERT.

6 Conclusion

In this paper, we focus on gender fairness in pre-trained language models. We first raised the concern and verified empirically that debiasing a pre-trained model with a second-phase pre-training approach leads to catastrophic forgetting issue. Then, we proposed GEnder Equality Prompt (GEEP) to alleviate gender bias in pre-trained language models without forgetting. In this approach, we freeze all model's parameters during second-phase pre-training and only update the embedding of the profession names as gender equality prompts. Results show that GEEP outperforms other models in gender bias reduction and downstream tasks without much forgetting.

REFERENCES

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.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

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*, 2020.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2232–2242, 2021.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115 (16):E3635–E3644, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/content/115/16/E3635.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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. 609–614, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

V Kocijan, O-M Camburu, A-M Cretu, Y Yordanov, P Blunsom, and T Lukasiewicz. Wikicrem: A large unsupervised corpus for coreference resolution. volume D19-1, pp. 4294–4303. Association for Computational Linguistics, 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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1030. URL https://aclanthology.org/N16-1030.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, 2018.

Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 777–789, 2012.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640, 2019.

Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *NeurIPS*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24, 2016.

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, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018.

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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853, 2018b.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized 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. 629–634, 2019.