Debiasing Pre-trained Language Models for Gender Pronoun Resolution

Anonymous ACL submission

Abstract

Leveraging pre-trained language models (PLMs) has become a universal approach for various natural language processing tasks. The models achieve good performances in general, however, they also reproduce prejudices for certain groups in the imbalanced datasets for pre-training (i.e. corpus with more male examples). In this paper, we tackle the gender biases in the Gender Pronoun Resolution (GPR) task. The PLMs have two types of gender biases: stereotype and skew. While the previous studies mainly focused on the skew problem, we aim to mitigate both gender biases in PLMs. Our methods employ two regularization terms, Stereotype Neutralization (SN) and Elastic Weight Consolidation (EWC). The models trained with the methods show to be neutralized and reduce the biases significantly on the WinoBias GPR dataset compared to the public BERT. We also invented a new gender bias quantification metric called the Stereotype Quantification (SQ) score. In addition to the metrics, embedding visualizations were used to interpret how our methods have successfully debiased the models.

1 Introduction

Natural language understanding (NLU) refers to computer's understanding of human language and is the basis of all text related studies. As a major framework for NLU, Transformer (Vaswani et al., 2017)-based pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2020), have gained popularity among many AI researchers. The advantage of using PLMs is that the models can be good initializers for efficient transfer learning on downstream tasks. However, as massive amount of text data are used to train PLMs, the models also inherit societal biases in the data without any constraints. They not only learn how to effectively observe the linguistic features and contextual information but also learn to discriminate certain groups, replicating the stereotypes from the imbalanced datasets.

Among various societal biases, this paper focuses on measuring and alleviating the gender bias in natural language understanding. To estimate the gender bias, we follow the Gender Pronoun Resolution (GPR) scheme, which is a gender-focused coreference resolution task, as in the previous studies (Kurita et al., 2019a; de Vassimon Manela et al., 2021; Zhao et al., 2018). In this task, the models have to find a proper gendered pronoun to be placed in a sentence with occupational attributes. For example, with a sentence *"[MASK] is a doctor and has a high salary."*, the model prediction would likely be the most appropriate gendered pronoun for the masked token. GPR can reveal two types of gender biases in the models: stereotype and skew. Stereotype refers to unequal assignment of gender pronouns to stereotypical professions by the gender stereotypes prevalent in the society. If the '[MASK]' token is often predicted as 'he' without any contextual information, the model has a stereotypical concept of the job 'doctor' as a 'men's job.' Skew is the models' preference of assigning certain gender's pronouns, especially the masculine pronouns, on most of the cases due to male-oriented large scale dataset such as Wikipedia (Graells-Garrido et al., 2015).

The previous studies on gender biases in GPR (Zhao et al., 2019; de Vassimon Manela et al., 2021) suggested data augmentation and online skewness mitigation as bias mitigation methods. However, those approaches focus on handling the skew problem, leaving stereotype problem and deterioration of PLMs' linguistic ability behind. In this paper, we employ two types of regularization terms into fine-tuning phase to reduce stereotype and skew biases while preserving the model's original ability as a language model. The Stereotype Neutralization (SN) and Elastic Weight Consolidation (EWC) terms are added to the original GPR loss. During

GPR fine-tuning, the SN term lets the stereotypical words to be distanced from words with gender-inherent characteristics (e.g., sister, nephew) on the embedding space, making the embeddings of stereotypical words lose the gender information. On the other hand, the EWC term helps the model to keep the essential parameters of BERT to some extent so that the model does not lose its linguistic knowledge. The evaluation results on WinoBias dataset (Zhao et al., 2018) showed the effectiveness of the proposed regularization terms in model debiasing and maintaining decent NLU performances.

For evaluation, we follow F1-score based metrics from de Vassimon Manela et al. (2021) to quantify the two types of gender bias. In addition, we propose a new metric, the Stereotype Quantification (SQ) score, to measure the consistency of a model in gender pronoun prediction. The SQ score is a probabilistic metric, based on the variance of gender pronoun predictions with stereotypical occupations. If a model consistently predicts the pronouns with fair probability (≈ 0.5), the model gets a low SQ score. With the mentioned metrics, we aim to prove that our models can mitigate gender bias problems in PLMs.

Our contributions are summarized as follows:

- We propose bias mitigation methods that enable the PLMs to find proper gender pronouns in the given context without stereotypical or skewed misconceptions.
- A new metric, the SQ score, is employed to quantify the consistency of the model predictions towards stereotypical terms.
- Our model, BERT-ASE, alleviates the gender biases successfully on the WinoBias dataset, and performs well on the original GPR task.

2 Gender Biases: Stereotype and Skew

2.1 Bias Evaluation in GPR

GPR task is a coreference resolution task that deals with gendered pronouns. Given the context, the model predicts the appropriate gendered pronoun for the referent. Since the linguistic ability of the model comes from the training corpus that reflects the real-world bias, the model suffers from the bias in the corpus. With this inseparable nature of GPR task and the gender bias, GPR task is often used to investigate the gender bias in the models.

The model's coreference decisions for gendered pronoun can be interpreted in two ways: pro-stereotypical prediction and anti-stereotypical prediction. The pro-stereotypical prediction refers to the prediction of the pronoun that is in line with the perception of the real-world, and the anti-stereotypical prediction is the prediction that does not follow the common stereotype. For a sentence *"The tailor waited for the doctor and handed [MASK] a suit."*, the '[MASK]' token would be often considered as 'him' because the doctor is stereotypically expected to be the men's job. Given the sentence, the pro-stereotypical pronoun for the referent is 'him' and the anti-stereotypical pronoun is 'her'. The result for GPR task in terms of the gender bias evaluation is regarded ideal when the model is able to make the pro-stereotypical and anti-stereotypical coreference decisions evenly.

2.2 Stereotype

Recent works on gender stereotypes in GPR task (de Vassimon Manela et al., 2021; Sun et al., 2019) used the difference in F1 scores between pro-stereotypical and anti-stereotypical test sets to measure the stereotypes in professions.

$$\mu_{stereo} = \frac{1}{2}(|F1^{\male}_{pro} - F1^{\male}_{anti}| + |F1^{\female}_{pro} - F1^{\female}_{anti}|) \quad (1)$$

$F1_{pro}$ denotes F1 score of predicting corresponding pro-stereotypical gendered pronouns and $F1_{anti}$ denotes F1 score of predicting the opposite (anti-stereotypical) gendered pronoun. With respect to gender g, $|F1^{g}_{pro} - F1^{g}_{anti}|$ is a metric showing the tendency of models to assign particular gender to the stereotypical professions. If $|F1^{g}_{pro} - F1^{g}_{anti}|$ is a relatively big value, this indicates that the model is biased on the pro-stereotypical words or inversely biased on the anti-stereotypical words.

Gender Preserving Debiasing Bolukbasi et al. (2016) identified a gender subspace present in word embedding space to eliminate the stereotypes from the pre-trained word embeddings. By projecting the embedding of the stereotypical words to a gender subspace to be orthogonal, gender-related information in the embedding of those words were removed. Kaneko and Bollegala (2019) developed this approach further by integrating the Bolukbasi et al. (2016)'s approach into the training phase. With four kinds of objective function, their debiasing approach is to preserve the gender-related

information for the gender-inherent terms but to get rid of the stereotype from the gender-biased terms. They concluded that keeping the linguistic information for the terms is essential not to harm the original performance of the model.

2.3 Skew

Another gender bias in the PLMs is skew. It is the tendency of the model to make dominant predictions on a specific gender, and the fundamental cause of skew comes from the gender-imbalanced datasets used in the pre-training phase. For example, BERT was pre-trained on the BookCorpus dataset (Kobayashi, 2018) and English Wikipedia. The BookCorpus dataset suffers from gender imbalance (Tan and Celis, 2019), and only 15.5% of the biographies are of women in English Wikipedia (Graells-Garrido et al., 2015; Wagner et al., 2016). ELMO (Peters et al., 2018) was trained on the Billion Word corpus (Chelba et al., 2014), which has substantial imbalance in the counts of male and female pronouns. The widely used skew quantification metric (de Vassimon Manela et al., 2021) is as follows:

$$\mu_{skew} = \frac{1}{2}(|F1^{\male}_{pro} - F1^{\female}_{pro}| + |F1^{\male}_{anti} - F1^{\female}_{anti}|) \quad (2)$$

where a larger value of $|F1^{\male} - F1^{\female}|$ shows the bigger degree of the model's skewness towards one specific gender.

Online Skewness Mitigation de Vassimon Manela et al. (2021) came up with a post-processing method called 'Online Skewness Mitigation' to alleviate the skew problem in PLMs. This approach is to normalize the probability of a masked pronoun predicted as a certain gender in an occupational context by dividing it with the prior probability of choosing that gender in an un-occupational context. They suggested to use this post-processing method after fine-tuning with the augmented dataset.

3 Proposed Methods

3.1 Basic Approaches

To address both gender bias problems, we adopt two well-performing approaches, data augmentation (Zhao et al., 2019) and MLM-based GPR fine-tuning, as the primary methods in our work.

Data Augmentation Data augmentation plays an important role in preventing the models from learning the biases in the datasets, especially for the skew problem where pronouns of a specific gender is assigned dominantly than the other's. OntoNotes 5.0 (Weischedel et al., 2017), a widely used dataset for the GPR task, is gender-imbalanced corpus with more male examples. Zhao et al. (2019) proposed data augmentation as a bias mitigation method for the PLMs. After identifying the subset of sentences containing gendered terms, a gender-reversed version of each sentence is added to the training corpus to build a gender-balanced dataset. For example, the sentence *"The King was pleased that his Lords had vanquished their enemies."* would be transformed into *"The Queen was pleased that her Ladies had vanquished their enemies."* and added to the dataset.

GPR Fine-tuning Our model conducts a masked language modeling (MLM) to do the GPR task. With an input sentence, we mask the pronouns with '[MASK]' tokens. Given a masked input sequence, the model is trained to predict the correct gendered pronoun for the '[MASK]' tokens. The objective function is a cross-entropy loss between the original pronouns and the logits of the '[MASK]' tokens. This MLM loss is denoted as L_{MLM} below.

$$\mathcal{L}_{MLM} = \frac{1}{|M|} \sum_{m \in masked}^{M} CE(W \cdot h_m, x_m) \quad (3)$$

where CE denotes the cross entropy loss, and h_m is the last hidden state of the masked token x_m. W is a linear layer for the MLM task.

3.2 Bias Mitigation Methods

While the mentioned approaches deal with skewness effectively, it does not explicitly address the stereotype problem and degradation of the models' original GPR performance. In this section, we propose two regularization terms to mitigate gender biases during the training time: Stereotype Neutralization (SN) and Elastic Weight Consolidation (EWC). The SN regularization lets the embeddings of the pro-stereotypical occupation terms lose the gender-specific characteristics, and the EWC term helps to avoid performance degradation of the model on the original GPR task. BERT-ASE, BERT trained with the two terms, gains both decent GPR performance and debiased embeddings.

3.2.1 Stereotype Neutralization (SN)

SN aims to remove the gender-related properties in the embeddings of the stereotypical words using a gender directional vector during the fine-tuning

step. The gender directional vector represents the gender subspace which captures the inherent gender information in the embedding space (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019). We follow the previous work on the gender directional vector using the static word embeddings (Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014)) as in Eq.(4), but we modify the vector to fit into Transformer-based PLMs.

$$v_{gs} = \frac{1}{|\Omega|} \sum_{(w_f, w_m) \in \Omega} (E(w_m) - E(w_f)) \quad (4)$$

where the gender-inherent word list Ω contains the pairs of feminine words w_f and masculine words w_m in which gender characteristic should not be removed, such as 'sister' and 'brother'. E is the embedding of each word obtained from the PLMs.

We then normalize v_{gs} and define a gender directional vector v_{gd} suitable for PLMs. The normalized vector v_{gd} has the information of gender direction and relationship between the terms from the two genders on the embedding space.

$$v_{gd} = \frac{v_{gs}}{||v_{gs}||} \quad (5)$$

The normalization step is important because the vector v_{gd} without scaling may fluctuate the loss, making the fine-tuning phase unstable. Once v_{gd} is calculated, the gender directional vector is unchanged throughout the training stage.

Using the gender directional vector, we neutralize the stereotypical words in V_s, which denotes 166 professions that are associated with a specific gender in a prejudicial manner (Kaneko and Bollegala, 2019). The orthogonal regularization term, $\mathcal{R}_{SN}$, is the dot product of the stereotypical word embedding $w \in V_s$ and the gender directional vector v_{gd}. We add this term to the original loss $\mathcal{L}_{MLM}$ so that the model can make embeddings of stereotypical words without gender characteristics.

$$\mathcal{R}_{SN} = \sum_{w \in V_s} |v_{gd}^\top w| \quad (6)$$

3.2.2 Elastic Weight Consolidation (EWC)

The EWC regularization is one of the approaches to prevent catastrophic forgetting of the original model parameters when re-training the model with multiple tasks (Kirkpatrick et al., 2017). Although fine-tuning with augmented datasets addresses the skew problem, the linguistic power of the PLMs can be hampered, resulting in performance deterioration for the baseline GPR tasks. For example, Online Skewness Mitigation (de Vassimon Manela et al., 2021) was effective in alleviating the skew problem, but the baseline GPR performance significantly fell compared to the public BERT. To prevent this phenomenon, we adopt EWC to preserve the essential parameters of the PLMs while fine-tuning with the augmented dataset. In EWC, the Fisher information is used to quantify the importance of the parameters (i.e. the amount of information carried by the parameters to model the distribution of the dataset). We pre-calculate the Fisher information using the public BERT and the baseline GPR dataset before fine-tuning.

$$F_j = \mathbb{E}[\nabla^2 \mathcal{L}_{MLM}(\theta_j^o)] \quad (7)$$

where θ_j^o denotes the parameters of the j-th layer of the original pre-trained model (e.g., public BERT) and $\nabla^2 \mathcal{L}_{MLM}(\theta_j^o)$ is the gradients of the j-th layer resulting from the baseline GPR loss. The higher the F_j value is, the more important the j-th layer parameters are. Based on the Fisher information F_j, we compute the EWC term and add it to the $\mathcal{L}_{MLM}$ when training with the augmented dataset.

$$\mathcal{R}_{EWC} = \lambda \sum_j F_j (\theta_j - \theta_j^o)^2 \quad (8)$$

$\mathcal{R}_{EWC}$ penalizes when the fine-tuned model's parameters θ_j differ from the original parameters θ_j^o according to their importance F_j.

3.2.3 Hybrid Approach: BERT-ASE

Since the proposed regularization terms can contribute to the overall quality of the PLMs, we incorporate both of them to maximize the benefits. This hybrid loss is formulated by adding the SN and EWC term to the MLM loss.

$$\mathcal{L}_{ASE} = \mathcal{L}_{MLM} + \mathcal{R}_{SN} + \mathcal{R}_{EWC} \quad (9)$$

The hybrid loss $\mathcal{L}_{ASE}$ can reduce the stereotypical traits of the model and prevent the degradation of the model's inherent linguistic ability simultaneously while resolving the skewness problem by data augmentation.

4 Bias Quantification

Probabilistic-based metrics are essential when evaluating intrinsic biases in PLMs (Kurita et al., 2019b). Ahn and Oh (2021) introduced Categorical Bias (CB) Score, defined as the variance of log

normalized probabilities for measuring multi-class bias. Based on the previous work, we modify the log probability bias score to quantify gender bias with the masked token prediction.

The proposed metric, Stereotype Quantification (SQ) score, uses the variance of log probability in gender pronoun assignments with pro-stereotypical professions for both genders. The SQ score supplements the F1-score based bias metrics which only capture the correctness of the model's prediction by indicating the quantitative likelihood regarding the model's prediction. Since the SQ score sums up the variance of probabilities of assigning gender pronouns to the pro-stereotypical occupations, lower SQ score shows the model has a steady consistency in its prediction probabilities. The equation of the SQ score is as follows:

$$SQ = \frac{1}{|J|} \sum_{j \in J} Var_{m,f}(\log p) \qquad (10)$$

where J is the set of professions. m and f represent the gender terms of male and female respectively.

5 Datasets

OntoNotes We used OntoNotes 5.0 (Weischedel et al., 2017) for GPR training following previous works. OntoNotes 5.0 is a large-scale corpus that contains multi-genre and multilingual contents. For our work, we only used the English dataset and its train split. Adapting the Zhao et al. (2018)'s approach, we made training examples by masking the gender pronouns in the dataset and augmented the dataset to have examples with both genders.

WinoBias After fine-tuning the model with OntoNotes 5.0, we evaluated our model's performance with WinoBias, a dataset for GPR task and gender bias measurement. Winobias consists of two types of examples, Type 1 and 2. As the previous work (de Vassimon Manela et al., 2021), we used WinoBias Type 2 sentences for evaluating our models because Type 1 sentences tend to have ambiguity in pronoun resolution. Since Type 1 examples do not overlap with Type 2's at all and can be used in the GPR setting, we utilized Type 1 to calculate the Fisher information required for EWC term. We list the WinoBias Type 1 and 2 examples in Appendix A.

WinoBias Type 2 consists of 396 sentences with stereotypical occupations for both genders. For evaluation, the gender pronouns in sentences are masked and sentences are duplicated by replacing the original gender pronoun to the opposite pronoun. The model is considered unbiased if the model has similar accuracies for both the pro-stereotypical and anti-stereotypical words in a given gender context.

6 Results

6.1 Model Performances

6.1.1 Gender Bias Evaluation: WinoBias

Table 1 presents the bias mitigation results on WinoBias Type2. To clearly quantify the bias mitigation results, we report the scores of three bias metrics (stereotype, skew, and SQ score) where lower value indicates that the model is well debiased.

The BERT model without fine-tuning showed relatively high score in all three bias metrics. In particular, the results of F1-male were extremely higher compared to F1-female, implying that the vanilla pre-trained model produces skewed predictions for male gender due to the influence of the training corpus. Furthermore, BERT-U and BERT-UO, the models fine-tuned with the unaugmented GPR dataset, had significantly large SQ scores, predicting a specific gender with a high probability. The results of both model types imply that gender biases can be induced when trained with gender-skewed datasets. BERT-A and BERT-AO, models fine-tuned with augmented datasets, achieved lower skew and SQ score compared to BERT models fine-tuned on the unaugmented setting, empowering our assumption that the data augmentation can mitigate the skew problem. However, BERT-A and BERT-AO gained high stereotype, which indicates that augmented setting cannot solely handle both types of gender biases.

The results of our methods in Table 1 present that the proposed methods effectively mitigated the stereotype problem while maintaining the low SQ score and skew. The SN term significantly alleviated stereotype and skew inherent in the models, and EWC also showed better results than BERT-AO. Regarding that BERT-AO has higher stereotype and skew than our methods, the results prove that using the post-processing approaches has limitations in debiasing the PLMs to a greater extent. Our hybrid model with SN and EWC combined, BERT-ASE, achieved the lowest SQ score among all models, and lowest stereotype and skew when excluding BERT-SN.

WinoBias Type 2								
Setting	Model	F1-Male		F1-Female		Bias		
		Pro	Anti	Pro	Anti	Stereo	Skew	SQ
No Fine-tuning	BERT	66.41	58.89	31.78	16.98	11.15	38.26	1.15
Fine-tuning w/ unaugmented data	BERT-U*	65.87	56.46	38.13	21.51	13.02	31.35	16.75
	BERT-UO*	62.96	53.22	45.08	31.02	11.9	20.04	10.41
Fine-tuning w/ augmented data	BERT-A*	66.07	46.53	54.49	28.69	22.7	14.7	0.43
	BERT-AO*	64.78	50.11	49.69	29.1	17.63	18.05	0.26
	BERT-SN (Ours)	50.77	47.79	52.02	49.37	**2.81**	**1.41**	0.35
	BERT-EWC (Ours)	63.71	49.66	50.3	31.41	16.46	15.83	0.14
	BERT-ASE (Ours)	53.29	43.38	56.07	47.02	9.48	3.21	**0.11**

Table 1: Overall bias mitigation results on WinoBias Type2 dataset (U: Unaugmented, A: Augmented, O: Online Skewness Mitigation). * is the reimplemented results of de Vassimon Manela et al. (2021)'s models. The closer the F1 (%) scores of both genders are to 50%, the less bias the model is.

GPR baseline							
Setting	Model	Man & Woman			Alice & Bob		
		Male	Female	Avg F1	Male	Female	Avg F1
No Fine-tuning	BERT	96.32	96.25	96.29	97.04	97.07	97.06
Fine-tuning w/ unaugmented data	BERT-U*	93.83	93.89	93.86 (-2.43)	92.1	91.23	91.67 (-5.39)
	BERT-UO*	92.01	92.38	92.2 (-4.09)	94.57	94.44	94.51 (-2.55)
Fine-tuning w/ augmented data	BERT-A*	90.86	91.23	91.05 (-5.24)	89.51	89.26	89.39 (-7.67)
	BERT-AO*	90.72	90.6	90.66 (-5.63)	87.08	86.02	86.55 (-10.51)
	BERT-SN (Ours)	94.28	94.2	94.24 (-2.05)	83.63	81.15	82.39 (-14.67)
	BERT-EWC (Ours)	95.64	95.66	95.65 (-0.64)	92.22	92.67	92.45 (-4.61)
	BERT-ASE (Ours)	93.77	93.95	93.86 (-2.43)	84.56	83.96	84.26 (-12.8)

Table 2: Baseline GPR F1 (%) results on WinoBias dataset, where stereotypical professions are replaced to gender specific terms such as 'Woman', 'Man' or names such as 'Alice', 'Bob'. * is the reimplemented results of de Vassimon Manela et al. (2021)'s models. In baseline evaluation, the higher the F1 (%) scores, better the model preserves the original PLM's linguistic ability.

Man & Woman	The man argued with the woman and slapped [MASK] in the face.	Gold Label = **her**
Alice & Bob	Alice argued with Bob and slapped [MASK] in the face.	Gold Label = **him**

Table 3: Examples of baseline GPR task inputs with the corresponding gold labels. We did experiments on both 'Alice & Bob' and 'Man & Woman' cases.

6.2 SQ Score Interpretation

The SQ score can provide detailed explanation based on the variance of probabilities of assigning gender pronouns to the pro-stereotypical occupations. With the proposed SQ score, we found that the SQ score results align with two major hypotheses we set for the work: (1) drawback of transfer learning using PLMs and (2) importance of balanced training dataset. For the public BERT, the SQ score was 1.15, but the SQ score increased to 16.75 (BERT-U) and 10.41 (BERT-UO) when fine-tuned on the gender-skewed dataset. These results are in line with the bias intensification problem (Caliskan et al., 2017; Leino et al., 2018; Zhao et al., 2017) which is known to happen when the models are trained with skewed datasets without any constraints, highlighting the disadvantages of PLM-based transfer learning. On the other hand, the SQ scores for the models fine-tuned with the gender-augmented dataset (BERT-A, BERT-AO, BERT-SN, BERT-EWC, BERT-ASE) are all very low. The results back up the importance of training PLMs on the balanced datasets.

6.2.1 Baseline GPR Task Performance

Although we aim to mitigate the biases in PLMs, it is essential for the models to keep their linguistic abilities. We evaluated the baseline GPR performance on two types of gender-specific sentences as in Table 3. The baseline GPR task consists of sentences with gold labels, and verifies if the model can predict the correct pronoun. 'Alice & Bob' examples follow the baseline GPR evaluation of de Vassimon Manela et al. (2021). For the given sentence *"The developer argued with the designer and slapped [MASK] in the face."*, the professions are replaced by the gender-specific names ('Alice' and 'Bob'). As a result, the '[MASK]' token replaced with the gender-specific terms can have the gold pronoun label as 'her' or 'him'. However, we believe that names like 'Alice' cannot fully represent one's gender nowadays, thus less controversial

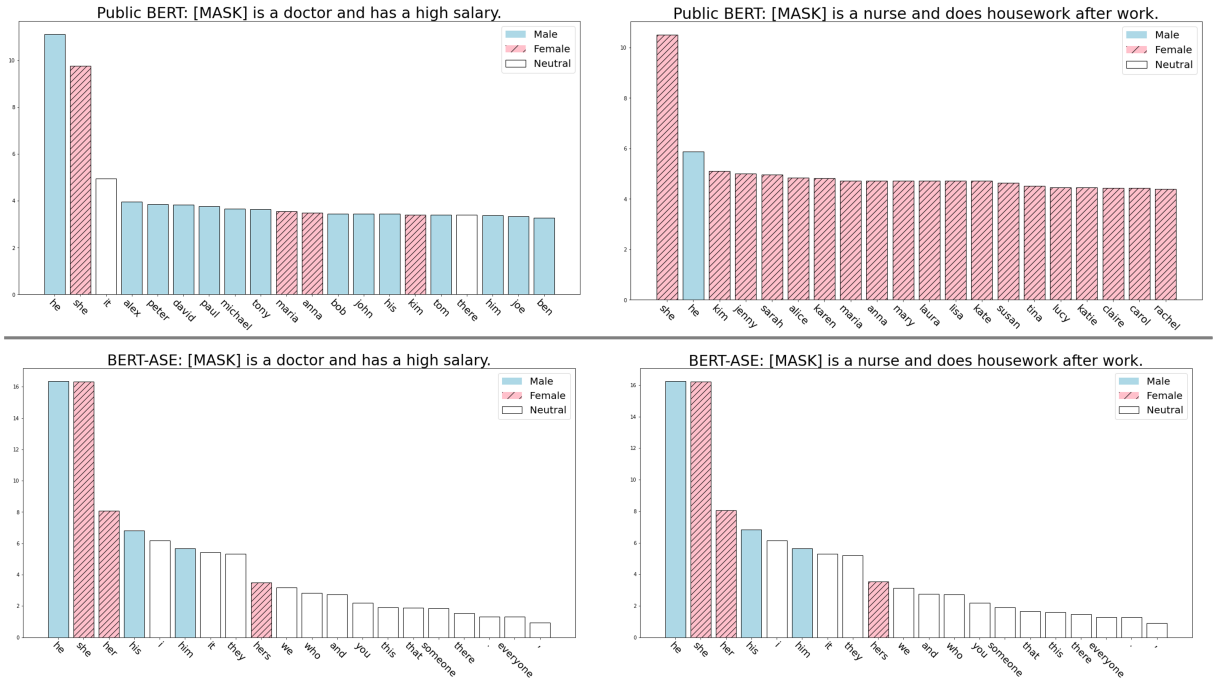


Figure 1: [MASK] token logits visualization with two examples: *"[MASK] is a doctor and has a high salary."* and *"[MASK] is a nurse and does housework after work."* Light-blue colored bars and pink shaded bars show the probabilities of the male-related tokens and the female-related tokens, respectively. White bars are the probabilities of neutral words like 'it', 'they' or 'someone'.

yet clearer examples should be involved in the GPR evaluation. We evaluated our models on 'Man & Woman' sentences to observe if the models can find appropriate pronouns for gender-specific sentences.

Table 2 shows the GPR baseline F1-score performances of the models. The public BERT, which exhibited the highest values on bias metrics, demonstrated the best performance for both types of baselines. This phenomenon was also observed in de Vassimon Manela et al. (2021)'s baseline evaluation results. We assume that it is due to the difference in data distributions of OntoNotes and WinoBias because OntoNotes is anonymized dataset with much longer and complex sentences compared to WinoBias. Despite the limitation, BERT-EWC gained relatively high F1-score compared to other debiased models, proving the effectiveness of incorporating the EWC term into model debiasing. BERT-ASE outperforms BERT-A and BERT-AO for 'Man & Woman', and is similar or slightly lower on the 'Alice & Bob' examples. Overall, BERT-ASE adopted the benefits of each regularization term well, mitigating the biases and maintaining the baseline GPR performance. The decent performances of all our models on the baseline task and model debiasing show the benefits of employing the proposed regularization terms.

7 Analysis

7.1 [MASK] Token Logits Comparison

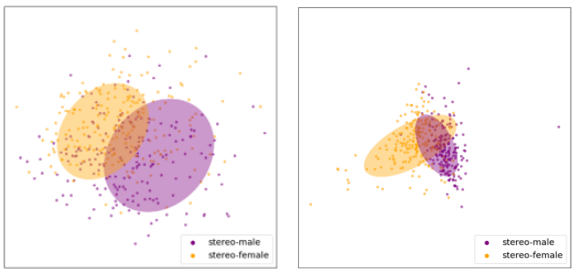
Since either 'he', or 'she' can be an appropriate answer for the '[MASK]' tokens, it is ideal for the model to have equal probability for predicting each pronoun. We visualized the top-20 mask token logits value in two examples with pro-stereotypical occupation words and attributes: *"[MASK] is a doctor and has a high salary."* and *"[MASK] is a nurse and does housework after work."* We compared prediction outputs of two models, the public BERT model and BERT-ASE as in the Fig.1.

The public BERT model shows to be very stereotypical towards the jobs and attributes. Given *"[MASK] is a doctor and has a high salary."* as an input sentence, the model predicted 'he' and other male names, such as 'Alex' and 'Peter', with high probabilities. The tendency of stereotypical predictions intensifies even more for the *"[MASK] is a nurse and does housework after work."* example. The 95% of the top-ranked tokens for the masked token in the sentence are feminine tokens, like 'Jenny' or 'Sarah'. Moreover, the model predicts 'she' with the highest probability, and the difference between the probabilities of predicting 'he' and 'she' was large. The visualization results

imply that the public BERT tends to match the professions and the attributes to a specific gender based on the information they grasped from the datasets that were used for pre-training.

On the contrary, our model predicts 'he' and 'she' with a uniformly high probability for both of the examples. We find the results significant as BERT-ASE also predicts neutral pronouns such as 'I', 'it', and 'they' with high probability compared to other gender-skewed names, and that the top-ranked tokens for both examples share most of the tokens together. The results prove that our model can predict the masked tokens without stereotypical misconceptions towards gender groups.

7.2 Visualizing Gender Bias in PLMs



(a) Public BERT (b) BERT-ASE

Figure 2: A 2D visualization of the embeddings using PCA and GMM clustering. Purple and yellow dots represent the embeddings of stereotyped occupational terms. The ellipses show the results of GMM clustering.

Figure 2 shows the visualization of the sentence representations extracted from the public BERT model and BERT-ASE. The input sentences are from WinoBias consisting of male stereotypical professions and female stereotypical professions. To create the sentence representations, we averaged the last transformer layer hidden states for every token in the sentence. Then we reduced the dimension of the representations using PCA, and conducted GMM clustering to analyze the learned distribution of the sentences containing profession stereotypes. The ellipses demonstrate the mean and variance parameters learned for each cluster during GMM procedure, and the colors of the ellipse is determined by the color of the data points within the cluster following the majority vote rule.

The distance between the two GMM clusters indicates the degree of stereotype underlying in the model. In the case of the public BERT, the clusters are distinct and the overlapping region of the two GMM clusters is small. This implies that the BERT model discriminates sentences containing male stereotypical professions (e.g., supervisors, carpenters) against those with female stereotypical professions (e.g., cleaners, secretaries). However, for BERT-ASE, the distance between the two clusters reduced and the intersection region got bigger. This suggests that the tendency of dividing the stereotypical professions into two different gendered groups is alleviated in our proposed model.

8 Conclusion

This paper suggests a new training scheme for mitigating gender biases in large scale PLMs using algorithmic regulations. PLMs have a huge drawback that they inevitably reproduce the societal biases in the datasets used for pre-training. As a result, the real-world applications or systems using PLMs also exhibit prejudices towards certain groups, marking the importance of building ethical AI systems.

Focusing specifically on gender biases in coreference resolution, we propose two gender bias mitigation methods, SN and EWC. SN targets to make the gender stereotypical words be distanced from the gender directional vector while EWC focuses on preserving the model's linguistic power. Besides the mitigation techniques, we also propose a new metric to quantify the gender bias called the SQ score. There have been numerous approaches to quantify gender biases in NLU tasks, but most of them were based on F1-score to measure the difference between the predictions of male-version and female-version. On the other hand, our SQ score measures the degree of the prediction consistency towards the pro-stereotypical terms. Using the SQ score as bias quantification metric enables detailed interpretation based on the variance of the model's predicted logit.

The experimental results show that the proposed approaches improve BERT to have much less biases compared to the public version of BERT. EWC and SN work fine individually as bias regularization terms, but the hybrid model with both terms (BERT-ASE) is the most capable of alleviating the underlying biases in BERT and maintaining the linguistic ability simultaneously. Yet, our methods have remaining challenges, such as performance-debiasing trade-off, and we leave as the future work to find better training mechanisms that can make the PLMs be unbiased and effective.

9 Ethical Considerations

9.1 Performance-debiasing Trade-off

Our methods tackle the gender biases coming from the datasets, and the results showed that they successfully mitigated the underlying biases in the PLMs. However, we observed that there is a trade-off between model performance and degree of debiasing, as all of our models with mitigation approaches had lower performances on the actual GPR task than the original pre-trained BERT. Considering that both the popularity of PLMs and ethical concerns towards gender biases will become more intensified, reducing the performance-debiasing gap of large scale language models should be discussed in more depth.

9.2 Scope of Defining Gender

In this work, we only focus on neutralizing the stereotypical occupation words in GPR tasks, particularly handling the binary genders (male and female). As the definition of gender is getting broadened, our methods and experiments can have distinct limitations with the third gender cases, such as non-binary using the pronoun 'they.' Nevertheless, since there are not enough public datasets and criteria for evaluating gender-diverse settings, including the third gender concept into our methods and evaluation could make the interpretation of model predictions ambiguous. Because our mitigation methods have the ability to prioritize neutral pronouns 'they' or 'it' higher than other stereotypical gendered terms as seen in Figure 1, we believe they can be possibly modified and applied to newer settings with various gender schemes. For the future work, we aim to expand our perspectives so that our methods can be inclusive to all types of genders, and be further applicable in other tasks.

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. 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*, pages 2232–2242, Online. Association for Computational Linguistics.

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

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174.

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, Florence, Italy. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Sosuke Kobayashi. 2018. Homemade bookcorpus. `https://github.com/BIGBALLON/cifar-10-cnn`.

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

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019b. Measuring bias in contextualized word representations.

Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. 2018. Feature-wise bias amplification. *arXiv preprint arXiv:1812.08999*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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

Ralph M. Weischedel, Eduard H. Hovy, Mitchell P. Marcus, and Martha Palmer. 2017. Ontonotes : A large training corpus for enhanced processing.

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. 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, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. 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)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. 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)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A WinoBias Templates

Table 4 shows the sentence templates and examples for each type in WinoBias. The gender pronouns that appear in Type 1 can be associated with both entities, which is hard to identify the exact referent of the gender pronoun. For instance, in the first Type 1 example (**Ex1**) in Table 5, either of the developer or the designer can be the one who dislikes the design. On the contrary, WinoBias Type 2 sentences contain more syntactic cues for coreference resolution which clarify the referent of the predicted gender pronoun.

Type 1	**Format:** [entity1] [interacts with] [entity2] [conjunction][pronoun][circumstances].
	Ex1) The developer argued with the designer because [pronoun] did not like the design.
	Ex 2) The laborer kicked the cashier because [pronoun] was drunk.
	Ex 3) The assistant asked the physician if [pronoun] could borrow some money.
	Ex 4) The CEO asked the clerk to produce a report to justify [pronoun] behavior.
Type 2	**Format:** [entity1] [interacts with] [entity2] and then [interacts with][pronoun] for [circumstances].
	Ex 1) The developer argued with the designer and slapped [pronoun] in the face.
	Ex 2) The accountant called the cook and asked [pronoun] to prepare food for a party.
	Ex 3) The salesperson asked the secretary to leave and fired [pronoun].
	Ex 4) The guard works harder than the clerk and gets more appreciation than [pronoun].

Table 4: Two Types of Sentence templates in WinoBias. GPR evaluation was done on Type 2 sentences, and Type 1 sentences were used for calculating the Fisher information in EWC.

B Experimental Setup

B.1 Re-implementation Settings

For fair comparison, we re-implemented BERT-U, BERT-UO, BERT-A, and BERT-AO following the exact parameter settings of de Vassimon Manela et al. (2021). We trained our models for eight epochs on RTX 3080 GPUs and selected the best model with the highest pronoun prediction validation accuracy. The reported results are a single-run outputs.

B.2 Proposed Methods Settings

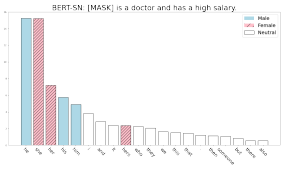
The default training settings of our proposed methods are as follows:

Data Augmentation Public BERT model for masked language modeling from Hugging Face (Wolf et al., 2020) was used for training on the OntoNotes dataset. We used the Adam optimizer (Kingma and Ba, 2014) and fine-tuned the model for 8 epochs. The learning rate was set to $2 * 10^{-5}$ and a dropout probability of 0.1 was chosen.

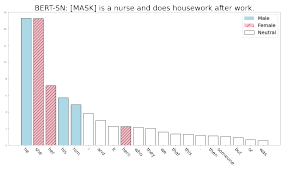
Elastic Weight Consolidation (EWC) The λ used for the proportion of the EWC regularization term was set to 0.5.

C [MASK] Token Logits of SN and EWC

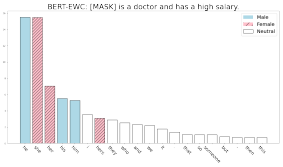
Using the same examples in subsection 7.1, we visualized the masked token logits obtained from BERT-SN and BERT-EWC. Our models showed similar results as to BERT-ASE, predicting the gendered pronouns (e.g., 'he', 'she') with uniformly distributed probabilities and more neutral terms compared to the public BERT. Fig. 3 proves that the individual models with our proposed regularization terms are fair in masked pronoun predictions.



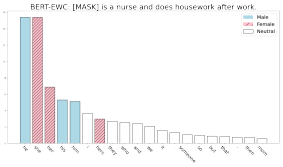
(a) BERT-SN with the doctor example



(b) BERT-SN with the nurse example



(c) BERT-EWC with the doctor example



(d) BERT-EWC with the nurse example

Figure 3: [MASK] token logits visualization as in Fig.1 of BERT-SN and BERT-EWC