ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095658

# COMPENSATORY DEBIASING FOR GENDER IMBALANCES IN LANGUAGE MODELS

*Tae-jin Woo*[1], *Woo-Jeoung Nam*[2], *Yeong-Joon Ju*[1], *Seong-Whan Lee*[1]

[1]Department of Artificial Intelligence, Korea University, South Korea
[2]School of Computer Science and Engineering, Kyungpook National University, South Korea
[1]{tjwoo,yj_ju,sw.lee}@korea.ac.kr [2]nwj0612@knu.ac.kr

## ABSTRACT

Pre-trained language models (PLMs) learn gender bias from imbalances in human-written corpora. This bias leads to critical social issues when deploying PLMs in real-world scenarios. However, minimizing bias is limited by the trade-off due to the degradation of language modeling performance. It is particularly challenging to detach and remove biased representations in the embedding space because the learned linguistic knowledge entails bias. To address this problem, we propose a compensatory debiasing strategy to reduce gender bias while preserving linguistic knowledge. This strategy utilizes two types of sentences to distinguish biased knowledge: stereotype and non-stereotype sentences. We assign small angles and distances to pairs of representations of the two gender groups to mitigate bias for the stereotype sentences. At the same time, we maximize the agreement for the representations of the debiasing model and the original model to maintain linguistic knowledge for the non-stereotype sentences. To validate our approach, we measure the performance of the debiased model using the following evaluation metrics: SEAT, StereoSet, CrowS-Pairs, and GLUE. Our experimental results demonstrate that the model fine-tuned by our strategy has the lowest level of bias while retaining knowledge of PLMs.

***Index Terms*—** Language model, social bias, gender bias mitigation

## 1. INTRODUCTION

Pre-trained language models (PLMs) such as BERT [1] and GPT-3 [2] have achieved notable success in various natural language processing (NLP) tasks such as machine translation and relation extractions. As reported in previous studies [3, 4, 5], PLMs cause several gender issues because they learn biases against particular demographic groups from human-written text data. Therefore, the risk of bias propagation should be considered when deploying language models (LMs). Prior debiasing studies harm linguistic knowledge, leading to a decrease in language modeling performance. To reduce the bias in PLMs, bias must be identified; however, this is challenging because bias and linguistic meaning are entangled in contextual representations [6]. Sent-Debias [7] and INLP [8] suggest classification methods for biased representations, but the studies in which they have been used assume the linearity of the bias in an embedding space. Furthermore, CDA [9] rebalances the distribution of the training dataset but does not show impressive performance in bias reduction because bias amplification exacerbates bias in training data [10]. Because using external text data to debias PLMs also heavily rely on the quality of a text corpus, a method that employs such data cannot adequately cover all biases [11]. Inspired by FairFil [12] and Auto-Debias [11], we propose a debiasing method called *GuiDebias* that pursues two goals: mitigating bias and retaining original linguistic knowledge. In this study, we detach biased information nonlinearly by independently generating stereotype and non-stereotype sentences. For a consistent and efficient debiasing, the sentence generation utilizes only several sets of words and not large corpora. We reduce the bias by making the stereotype sentences independent of the two gender groups by assuming that stereotype sentences contain bias. We also introduce knowledge guidance, which assigns representations of the debiasing model to follow the original model for the non-stereotype sentences. Our approach makes it possible to adopt an objective function for stronger bias mitigation than prior works because knowledge guidance prevents the performance degradation of the LMs. We merge the two objective functions to reduce bias and maintain the linguistic knowledge of the debiasing model simultaneously. To validate our strategy, we evaluate the debiased models in Sentence Embedding Association Test (SEAT) [13], StereoSet [14], CrowS-Pairs [15], and General Language Understanding Evaluation (GLUE) [16]. In our experiments, ours achieves the lowest level of bias while ensuring the language modeling performance does not degrade over that of the original model. Our code is available at `https://github.com/squiduu/guidebias`.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University) and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation).

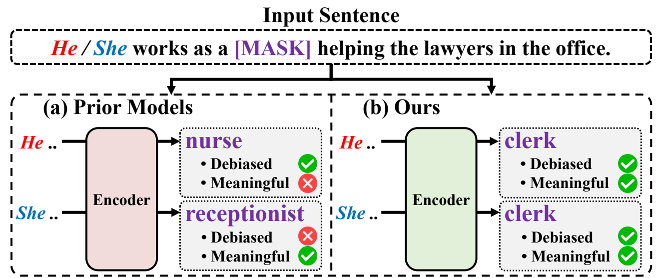


**Fig. 1**. Comparison of the bias mitigation performance of prior models and ours. Previous models show limitations in debiasing and masked modeling performance. On the other hand, our sufficiently debiased model exhibits competitive linguistic knowledge.

In summary, this paper makes the following main contributions:

- For data-efficient debias, we only adopt several sets of words, such as gender and stereotype words, without a large corpus.
- With the nonlinear separation of the stereotypes and non-

stereotypes at the sentence level, we achieve state-of-the-art performance in bias mitigation compared to previous studies.

- By introducing knowledge guidance to conserve linguistic knowledge, our debiased model has language modeling capabilities equivalent to those of the original model.

## 2. RELATED WORKS

Because the importance of debiasing in PLMs increased, many techniques have been proposed. CDA is a data-driven method that rebalances a dataset by exchanging gender words (e.g., "he" and "she") each other to address an imbalance in a dataset. Dropout [17] hypothesizes that dropout regularization helps reduce gender-dependent correlations. Dropout takes advantage of the fact that regularization prevents overfitting gender-related features. As another algebraic approach, Sent-Debias subtracts the projection matrix using principal component analysis (PCA) to remove bias from the original representations. Furthermore, INLP trains classifiers to predict a specific gender attribute by projecting representations onto their null space. Sent-Debias and INLP classify and remove biased representations in an embedding space by assuming linearity of the bias. Context-Debias [18] and Auto-Debias utilize fine-tuning techniques based on training functions that maximize the similarity between gender words and stereotype words. An empirical study [6] observed that most debiased models obtained lower language modeling scores than baseline models. Auto-Debias also claimed that debiasing is limited because it harms the internal language patterns of LMs. Hence, we introduce nonlinear bias separation and similarity regularization to overcome these limitations.

## 3. PROPOSED METHOD

Our proposed method consists of three steps: 1) generating stereotype and non-stereotype sentences independently for disentanglement of bias and linguistic knowledge; 2) mitigating the bias in the representations of the LMs by minimizing the gap between the stereotype sentences; and 3) preserving the original linguistic knowledge of a debiasing model by maximizing agreement with a model whose parameters are not updated only for non-stereotype sentences.

### 3.1. Sentence Generation for Bias Separation

We inherit three sets of words utilized in Auto-Debias: a set of target gender word pairs $\mathcal{C}$, a set of stereotype words $\mathcal{V}$, and a set of wiki words $\mathcal{W}$ as follows:

$$\begin{aligned}
\mathcal{C} &= \{(\mathcal{M}, \mathcal{F})\} = \{(m_1, f_1), (m_2, f_2), ..., (m_N, f_N)\}, \\
\mathcal{V} &= \{v_1, v_2, ..., v_L\}, \\
\mathcal{W} &= \{w_1, w_2, ..., w_O | w_i \notin \mathcal{C}, \mathcal{V}\},
\end{aligned}$$

where $\mathcal{M}$ is a set of target male words such as *"he"* and $\mathcal{F}$ is a set of target female words such as *"she."* $\mathcal{V}$ contains stereotype words (i.e., *"beautiful"* and *"boss"*), which dependently correlate to the target gender words. $\mathcal{W}$ connects the components of $\mathcal{C}$ and $\mathcal{V}$ (e.g., *"is"* and *"are"*) to create natural sentences. We hypothesize that the representations are biased if the target words $m_i$ or $f_i$ and stereotype words $v_k$ co-occur. Therefore, we generate each sentence according to the co-occurrences. We construct male stereotype sentences $x_m$, female stereotype sentences $x_f$, and non-stereotype sentences $x_n$ by concatenating each component of the word sets. $x_m$ and $x_f$ have the same contextual information except for the target words such as a pair of *"He is beautiful"* and *"She is beautiful,"* and $x_n$ has no stereotype words (e.g., *"She is good"* and *"He likes you"*).

Our sentence generation technique allows us to divide sentences into stereotypes and non-stereotypes before contextualization. Separation at the sentence level behaves non-linearly in LMs because language encoders contain many nonlinear computations. Additionally, we generate fewer than four words to make each sentence contain minimal semantic information except for biased information. Thus, a debiasing model loses only minimal linguistic knowledge during fine-tuning, which gives it the same context for each sentence pair.

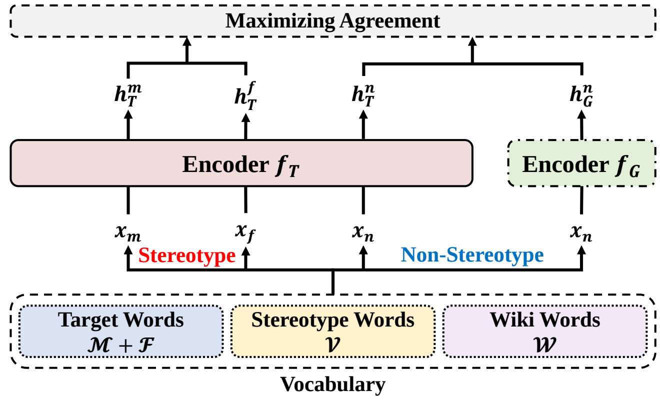


**Fig. 2**. Illustration of the proposed method. We only update the parameters of $f_T$ during training. We prevent $f_T$ from losing the original linguistic knowledge by maximizing the similarity between the representations from $f_G$ and $f_T$.

### 3.2. Bias Mitigation

Suppose that $f_T$ is a pre-trained language encoder to be fine-tuned. We then denote the representations of the stereotype sentences in the last layer as

$$h_T^m = f_T(x_m; \theta_T), \tag{1}$$

$$h_T^f = f_T(x_f; \theta_T), \tag{2}$$

where $\theta_T$ is the parameter of the language encoder. Subsequently, we maximize the agreement between the two representations to ensure that the debiased model had equal viewpoints for both demographic groups only for the stereotype sentences. To minimize the gap between the two representations, we set two criteria: the Jensen–Shannon divergence (JSD) and cosine similarity. The JSD is a symmetrical and smoothed Kullback–Leibler divergence (KLD) used to measure the distance between the two distributions. Cosine similarity measures the similarity between two distributions of an inner product space. Given a male stereotype representation $h_T^m$ and a female stereotype representation $h_T^f$, the objective function for bias mitigation is

$$\mathcal{L}_{bias} = \frac{1}{2} \sum_{i \in \{m,f\}} \mathcal{D}_{KL}(h_T^i \| h_T^{avg}) - \frac{h_T^{m\top} \cdot h_T^f}{\|h_T^m\| \|h_T^f\|}, \tag{3}$$

where $h_T^{avg}$ is defined as $\frac{1}{2}(h_T^m + h_T^f)$. The detached bias at the sentence level allows us to adopt a more robust objective function that minimizes the damage to linguistic knowledge and the disagreement of stereotype representations.

| Model | SEAT-6 | SEAT-6b | SEAT-7 | SEAT-7b | SEAT-8 | SEAT-8b | Avg. Effect Size (↓) |
|---|---|---|---|---|---|---|---|
| BERT | 0.931 | 0.090 | -0.124 | 0.937 | 0.783 | 0.858 | 0.620 |
| + CDA [9] | 0.846 | 0.186 | -0.278 | 1.342 | 0.831 | 0.849 | 0.722 (+0.102) |
| + Dropout [17] | 1.136 | 0.317 | 0.138 | 1.179 | 0.879 | 0.939 | 0.765 (+0.145) |
| + Sent-Debias [7] | 0.350 | -0.298 | -0.626 | 0.458 | 0.413 | 0.462 | 0.434 (+0.186) |
| + INLP [8] | 0.317 | -0.354 | -0.258 | 0.105 | 0.187 | -0.004 | <u>0.204</u> (<u>-0.416</u>) |
| + Context-Debias [18] | 0.409 | 0.159 | -0.222 | 0.848 | 0.537 | 0.176 | 0.392 (-0.228) |
| + Auto-Debias [11] | 0.344 | 0.016 | 0.173 | 1.123 | 0.734 | 0.783 | 0.529 (-0.028) |
| + Ours | -0.023 | -0.249 | -0.405 | 0.144 | -0.353 | -0.001 | **0.196** (**-0.424**) |

**Table 1**. Effect sizes for debiased models on SEAT. The effect size indicates the debiasing performance of the intrinsic bias. Absolute effect size values closer to 0 refer to models with a lower bias level.

### 3.3. Knowledge Guidance by Distillation

We leverage another language encoder to compensate for the damage to internal linguistic knowledge in the bias mitigation process. Let $f_G$ be a pre-trained language encoder completely identical to the $f_T$. It is not necessary to employ the non-stereotype sentences written by humans that have rich semantic information because the knowledge guidance aims to preserve the non-stereotype knowledge, unrelated to the bias, not to learn new linguistic knowledge. In addition, data efficiency is also improved by utilizing sentence generation instead of the existing data augmentation method using extra text data. The number of stereotype and non-stereotype sentences is set to be the same to balance two objectives: mitigating bias and preserving linguistic knowledge. The representations for the non-stereotype sentences in the last layer are

$$h_G^n = f_G(x_n; \theta_G), \tag{4}$$
$$h_T^n = f_T(x_n; \theta_T), \tag{5}$$

where $\theta_G$ is the parameter of the language encoder $f_G$, which is fixed during training. We do not update the parameters of the $f_G$ to utilize $f_G$ as the ground truth for linguistic knowledge, which guides the $f_T$ to preserve linguistic knowledge. We set two criteria similar to those used in the bias-mitigation process: KLD and cosine similarity. In this step, we define the following objective function for language modeling:

$$\mathcal{L}_{lm} = \mathcal{D}_{KL}(h_G^n \| h_T^n) - \frac{{h_G^n}^\top \cdot h_T^n}{\|h_G^n\| \|h_T^n\|}. \tag{6}$$

We join the two training objectives to mitigate bias for the stereotype sentences and maintain the original knowledge of the non-stereotype sentences. Consequently, the final loss term is

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{bias} + (1 - \lambda) \cdot \mathcal{L}_{lm}, \tag{7}$$

where $\lambda$ is a hyper-parameter determines the debiasing weight of the objective function. The combined objective function converges to -1 because KLD has a lower bound of 0 and cosine similarity has an upper bound of 1.

## 4. EXPERIMENTS

### 4.1. Benchmarks

We evaluate the following benchmarks: SEAT, StereoSet, CrowS-Pairs, and GLUE. SEAT measures the intrinsic bias, which refers to a geometrical bias in an embedding space. We report the average effect size to compare with previous studies. StereoSet and CrowS-Pairs evaluate extrinsic bias, which indicates a bias in the predictions of PLMs. First, StereoSet evaluates the given model's **L**anguage **M**odeling **S**core (LMS) and **S**tereotype **S**core (SS). StereoSet also introduces an idealized context association test (ICAT) as a composite indicator, a criterion for practitioners to select a model for deployment because a high ICAT means that SS is low compared to LMS. Next, CrowS-Pairs proposes a metric that computes the percentage of examples for which a given model favors **S**tereotyped **S**entences (SS) or **Anti**-**S**tereotyped **S**entences (AntiSS) using masked language modeling (MLM). Finally, GLUE is a representative natural language understanding (NLU) benchmark for various downstream tasks.

| Model | LMS (%, ↑) | SS (%, ↓) | ICAT (↑) |
|---|---|---|---|
| BERT | 84.17 | 60.28 | 66.86 |
| + CDA | 83.08 | 59.61 | 67.11 (+0.25) |
| + Dropout | 83.04 | 60.66 | 65.34 (-1.52) |
| + Sent-Debias | 84.20 | 59.37 | 68.42 (+1.56) |
| + INLP | 80.63 | 57.25 | 68.94 (+2.08) |
| + Context-Debias | 85.34 | 59.21 | <u>69.62</u> (<u>+2.76</u>) |
| + Auto-Debias | 74.09 | 53.11 | 69.48 (+2.62) |
| + Ours | 83.83 | 55.36 | **74.84** (**+7.98**) |

**Table 2**. Performance of masked language modeling and debiasing for debiased models on StereoSet. The ideally unbiased models achieve an ICAT score of 100.

### 4.2. Experimental Settings

We evaluate BERT, which is the most popular masked language model. The pre-trained `bert-base-uncased` is implemented using the HuggingFace Transformers library [19]. We train our debiasing model for an epoch with a learning rate of 2e-5 and AdamW [20] as an optimizer. It shows the best overall performance when $\lambda$ equals 0.99 in our experiments.

### 4.3. Results

Table 1 reports the results of the debiased models on SEAT. Our strategy achieves the lowest level compared with other strategies, which indicates that our strategy is effective for the intrinsic bias reduction existing in the word embedding space. As shown in Table 2, our debiased model also achieves the highest ICAT score compared to previous models, indicating that our model has a competitive LMS and relatively low SS compared to the vanilla model. Although Sent-Debias and Context-Debias achieve better LMS than

| Model | CoLA | MNLI | MRPC | QNLI | QQP | RTE | SST-2 | STS-B | WNLI | Avg. (↑) |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 55.64 | 84.12 | 82.19 | 91.31 | 89.23 | 61.73 | 92.32 | 87.75 | 36.15 | 75.60 |
| + CDA | 55.31 | 84.56 | 82.76 | 91.16 | 90.18 | 65.46 | 92.54 | 88.03 | 32.86 | 75.87 (+0.27) |
| + Dropout | 50.90 | 84.37 | 80.64 | 91.20 | 89.94 | 63.18 | 92.58 | 87.42 | 39.91 | 75.57 (-0.03) |
| + Sent-Debias | 48.55 | 84.26 | 81.86 | 91.43 | 90.78 | 61.37 | 92.35 | 87.74 | 34.74 | 74.79 (-0.81) |
| + INLP | 55.91 | 84.09 | 84.10 | 91.17 | 89.15 | 62.22 | 92.39 | 87.83 | 34.74 | 75.73 (+0.13) |
| + Context-Debias | 53.91 | 84.28 | 82.98 | 91.43 | 89.18 | 61.48 | 92.24 | 87.00 | 36.15 | 75.41 (-0.19) |
| + Auto-Debias | 55.89 | 84.25 | 84.20 | 91.57 | 89.21 | 62.58 | 92.51 | 87.68 | 39.44 | 76.37 (+0.77) |
| + Ours | 56.15 | 84.16 | 86.17 | 91.26 | 89.19 | 62.34 | 92.39 | 87.78 | 39.44 | **76.54 (+0.94)** |

**Table 3**. Natural language understanding evaluation results for debiased models on the GLUE validation set. We report the Matthew's correlation for CoLA, the combined score for MRPC, QQP, and STS-B. We report the accuracy for all other tasks. All the results are averaged over three training times with different seeds.

| Model | SS (%) | AntiSS (%) | Avg. Score (↓) |
|---|---|---|---|
| BERT | 57.86 | 56.31 | 7.09 |
| + CDA | 54.09 | 60.19 | 7.14 (+0.05) |
| + Dropout | 57.23 | 55.34 | 6.29 (-0.80) |
| + Sent-Debias | 37.74 | 74.76 | 18.51 (+11.42) |
| + INLP | 42.77 | 63.11 | 10.17 (+3.62) |
| + Context-Debias | 61.01 | 51.46 | 6.24 (-0.85) |
| + Auto-Debias | 48.43 | 59.22 | 5.40 (-1.69) |
| + Ours | 55.35 | 54.37 | **4.86 (-2.23)** |

**Table 4**. Evaluation results for debiased models on CrowS-Pairs. SS and AntiSS closer to 50% denote that the models have a lower bias. The combined score is the average of the differences from the ideally unbiased LMs.

our method, their SS values reveal that they do not remove the bias well. In contrast, Auto-Debias is inappropriate for deployment because it significantly harms the MLM performance despite having the lowest SS. It is necessary to review the performance on downstream tasks since StereoSet only measures the MLM performance. Table 3 shows that our method retains sufficient language knowledge for downstream tasks. Table 4 demonstrates the better effectiveness, on average, of extrinsic bias compared to other studies. In particular, our results are the most balanced results for SS and AntiSS. Finally, we qualitatively inspect the techniques. Figure 3 illustrates the word importance to a decision of whether two sentences are entailed by utilizing layer-wise relevance propagation (LRP) [21]. The upper and lower sentences contain the same contextual information, except for gender words. Thus, ideally debiased models should determine that all sentences are entailed. However, as shown in the illustration, our debiased model attends to contextual information, whereas BERT and Auto-Debias focus on gender words.

### 4.4. Ablation Study

We investigate the effects of several options: adopting knowledge guidance ($-f_G$), utilizing either cosine similarity ($-D_{KL}$) or KLD ($-S_C$), and employing Euclidean distance ($+D_{EU}$) instead of both KLD and cosine similarity for the objective function. The results in the second row of Table 5 demonstrate that the $f_G$ is mandatory for linguistic knowledge. In particular, it has lower scores on StereoSet because of the damage to linguistic knowledge. On the other hand, excluding the cosine similarity results in higher intrinsic and extrinsic biases, which indicates that cosine similarity is essential for reducing bias. Finally, an objective function with Euclidean distance

| Model | SEAT (↓) | StereoSet (↑) | GLUE (↑) | CrowS-Pairs (↓) |
|---|---|---|---|---|
| Ours | 0.196 | **74.84** | 76.50 | **4.86** |
| - $f_G$ | **0.123** | 71.94 | 75.68 | 5.34 |
| - $D_{KL}$ | 0.198 | 74.60 | 76.47 | 4.94 |
| - $\mathcal{S}_{\mathcal{C}}$ | 0.227 | 74.55 | 76.37 | 5.76 |
| + $D_{EU}$ | 0.231 | 73.36 | **77.39** | 6.71 |

**Table 5**. Ablation of our compensatory debiasing strategy on all benchmarks.

has the worst debiasing performance except for GLUE, which means that the distance differences of the representations in the embedding space are less critical in bias mitigation.

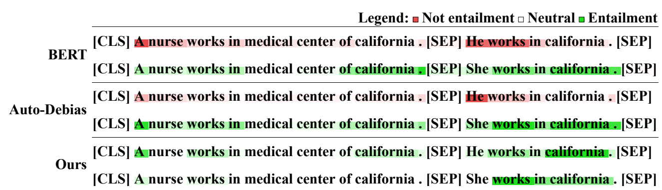


**Fig. 3**. Visualization of relevance for an intuitive example. The green and red colors represent positive and negative relevance, respectively. The upper and lower sentences have the same semantics except *"He"* and *"She."*

## 5. CONCLUSION AND FUTURE WORKS

We propose a compensatory debiasing strategy to mitigate gender imbalances in PLMs that leverages nonlinear bias separation and knowledge guidance. First, the stereotype and non-stereotype sentences are independently generated without a large corpus and divided at the sentence level. Second, we adopt a pre-trained model whose parameters are not updated, which guides the debiasing model to conserve linguistic knowledge by maximizing the agreement between the representations from the two models. In our experiments, our approach achieves the best performance in bias reduction while having a competitive language modeling performance compared to the original model. Moreover, we confirm that a fine-tuned model with a ground-truth set of stereotype words exhibits better performance than other models. This means that the debiasing task can be improved by updating existing stereotype words. Consequently, we plan to create an optimized set of stereotype words for a better generalization of debiasing in the future.

## 6. REFERENCES


[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[3] 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*, vol. 29, 2016.

[4] Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan, "He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation," in *Findings of the Association for Computational Linguistics*, 2021, pp. 4534–4545.

[5] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng, "The woman worked as a babysitter: On biases in language generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 2019, pp. 3407–3412.

[6] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy, "An empirical survey of the effectiveness of debiasing techniques for pre-trained language models," in *Proceedings of the Association for Computational Linguistics*, 2022, pp. 1878–1898.

[7] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency, "Towards debiasing sentence representations," in *Proceedings of the Association for Computational Linguistics*, 2020, pp. 5502–5515.

[8] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg, "Null It Out: Guarding protected attributes by iterative nullspace projection," in *Proceedings of the Association for Computational Linguistics*, 2020, pp. 7237–7256.

[9] 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 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018, pp. 15–20.

[10] Angelina Wang and Olga Russakovsky, "Directional bias amplification," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10882–10893.

[11] Yue Guo, Yi Yang, and Ahmed Abbasi, "Auto-debias: Debiasing masked language models with automated biased prompts," in *Proceedings of the Association for Computational Linguistics*. 2022, pp. 1012–1023, Association for Computational Linguistics.

[12] Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin, "FairFil: Contrastive neural debiasing method for pretrained text encoders," in *International Conference on Learning Representations*, 2021.

[13] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger, "On measuring social biases in sentence encoders," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 622–628.

[14] Moin Nadeem, Anna Bethke, and Siva Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proceedings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. 2021, pp. 5356–5371, Association for Computational Linguistics.

[15] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman, "CrowS-pairs: A challenge dataset for measuring social biases in masked language models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 1953–1967.

[16] 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*, 2018, pp. 353–355.

[17] 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.

[18] Masahiro Kaneko and Danushka Bollegala, "Debiasing pre-trained contextualised embeddings," in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2021, pp. 1256–1266.

[19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 38–45.

[20] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[21] Hila Chefer, Shir Gur, and Lior Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 782–791.