Debiasing isn't enough! – On the Effectiveness of Debiasing MLMs and their Social Biases in Downstream Tasks

Masahiro Kaneko[1]   **Danushka Bollegala**[2,3*]   **Naoaki Okazaki**[1]
[1]Tokyo Institute of Technology   [2]University of Liverpool   [3]Amazon
masahiro.kaneko@nlp.c.titech.ac.jp
danushka@liverpool.ac.uk   okazaki@c.titech.ac.jp

Abstract

We study the relationship between task-agnostic intrinsic and task-specific extrinsic social bias evaluation measures for Masked Language Models (MLMs), and find that there exists only a weak correlation between these two types of evaluation measures. Moreover, we find that MLMs debiased using different methods still re-learn social biases during fine-tuning on downstream tasks. We identify the social biases in both training instances as well as their assigned labels as reasons for the discrepancy between intrinsic and extrinsic bias evaluation measurements. Overall, our findings highlight the limitations of existing MLM bias evaluation measures and raise concerns on the deployment of MLMs in downstream applications using those measures.

1 Introduction

Text representations produced by MLMs have revolutionised NLP by improving the performance of numerous downstream applications (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; Yang et al., 2019). Unfortunately, large-scale pretrained MLMs demonstrate worrying levels of social biases when applied to downstream tasks (May et al., 2019; Nadeem et al., 2021; Kaneko and Bollegala, 2021a; Kaneko et al., 2022b). Given that real-world Natural Language Processing (NLP) systems such as machine translation systems, dialogue systems, etc. are used by millions of users world-wide (Hovy and Spruit, 2016), it remains an important responsibility to accurately evaluate the social biases in MLMs prior to deployment.

Two types of social bias evaluation measures have been proposed for MLMs in prior work (Goldfarb-Tarrant et al., 2021; Cao et al.,

2022): (a) task-agnostic *intrinsic* evaluation measures (Nangia et al., 2020; Nadeem et al., 2021; Kaneko and Bollegala, 2022) that use the likelihood scores assigned by an MLM under evaluation for sentences representing various social biases, and (b) task-specific *extrinsic* evaluation measures that use the data from downstream NLP tasks such as predicting the occupation of a person from their biographies (De-Arteaga et al., 2019; Bartl et al., 2020). In contrast to intrinsic measures that directly probe into the biases in MLMs, extrinsic measures holistically evaluate an NLP system that uses an MLM.

Much prior work has decoupled the intrinsic and extrinsic bias evaluations for simplicity, and have assumed that an MLM unbiased according to an intrinsic measure will remain to be so under extrinsic evaluations too. However, this unverified hypothesis begs the following question: *Can we reliably predict the social biases of an MLM when it is applied to a particular downstream task using only intrinsic evaluation measures?* To answer this question, we conduct a comprehensive study using 7 pretrained MLMs and their debiased versions using three different debiasing methods. However, only weak correlations are found between intrinsic and extrinsic bias evaluation measures. This is a worrying proposition because intrinsic measures are often used to decide whether an MLM should be deployed in a downstream application.

We further investigate on why MLMs learn social biases during downstream task fine-tuning and identify two sources: *instance-related* biases and *label-related* biases. The training instances for downstream tasks contain unfair associations that are learnt by the masked language modelling objective, which we name instance-related biases. We find that debiased MLMs gradually learn such biases during fine-tuning, leading to disagreements between intrinsic and extrinsic evaluation measures. Moreover, the labels assigned to downstream train-

---

ing instances can also represent biased ratings, which are also learnt by the MLMs when their parameters are updated in an end-to-end manner when predicting task labels. Overall, we find weak correlations between intrinsic and extrinsic measures even after the models are fine-tuned on downstream task data. Based on our findings, we recommend that intrinsic bias evaluation measure alone must not be used to determine whether an MLM should be deployed in a downstream NLP application. The code and data used in this paper will be publicly released to facilitate reproduction of results upon paper acceptance.

2 Related Work

Social biases have been reported in models trained for numerous downstream tasks. Kiritchenko and Mohammad (2018) examined 219 automatic sentiment analysis systems that participated in the SemEval-2018 Task 1 (Mohammad et al., 2018) and found that many systems show statistically significant gender or racial biases. Díaz et al. (2019) investigated the age-related biases in sentiment classification and found that across 15 sentiment analysis tools, sentences containing adjectives describing the youth were 66% more likely to be scored positively than those describing the elderly. The problem is also observed in MLMs trained for many languages (Kaneko et al., 2022b).

Numerous bias mitigation approaches have been proposed for MLMs such as fine-tuning (Kaneko and Bollegala, 2021a; Lauscher et al., 2021), counterfactual data augmentation (Hall Maudslay et al., 2019; Zmigrod et al., 2019) and parameter dropout (Webster et al., 2020). Liang et al. (2020a) proposed a method to debias the sentence embeddings created from BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) inspired by hard debiasing (Bolukbasi et al., 2016), a method originally proposed for static word embeddings. However, not all biases are adequately mitigated by the current proposals (Gonen and Goldberg, 2019). Our focus in this paper is the evaluation and not debiasing methods.

Intrinsic bias evaluation measures (Nangia et al., 2020; Nadeem et al., 2021; Kaneko and Bollegala, 2022) evaluate the social biases in a given MLM standalone, independently of any downstream applications. Pseudo log-likelihood scores assigned to stereotypical vs. antistereotypical examples (either manually written or automatically generated

using templates) have been used to evaluate the social biases in an MLM as further detailed in § 3. Considering that MLMs are used to represent input texts in various downstream tasks, several prior work have argued that their social biases must be evaluated with respect to those tasks (De-Arteaga et al., 2019; Bartl et al., 2020; Webster et al., 2020). However, decoupling tasks and MLMs makes the bias evaluation simpler and task-independent, which is attractive because MLMs are designed as generic text representations for a wide-range of downstream tasks. We further discuss extrinsic bias evaluation measures in § 3.

Goldfarb-Tarrant et al. (2021) studied social biases in static word embeddings using the Word Embedding Association Test (WEAT) as an intrinsic measure and coreference detection and hate speech detection as extrinsic evaluation tasks. They found the correlations between intrinsic and extrinsic measures to be weak or negative. Although we share the motivation with Goldfarb-Tarrant et al. (2021), our focus is contextualised embeddings produced by MLMs, which have outperformed static word embeddings in numerous tasks, hence more widely used in real-world NLP applications than static word embeddings.

Cao et al. (2022) studied the correlation between intrinsic bias scores (i.e. Contextualised Embedding Association Test (Guo and Caliskan, 2021, CEAT) and Increased Log Probability Score (Kurita et al., 2019, ILPS)) and extrinsic bias scores for three tasks: toxicity (Jigsaw, 2019), hate-speech (Mathew et al., 2021) and sentiment (Dhamala et al., 2021) detection. Similar to Goldfarb-Tarrant et al. (2021) they also found weak correlations. As reasons for this lack of correlation they highlight misalignment between the metrics such as the notion of bias, protected groups, and noise in the evaluation datasets. Our contributions are complementary to theirs because we consider different intrinsic and extrinsic measures, not covered by them. Moreover, our extrinsic tasks are directly related to social biases. We observe weak correlations between intrinsic and extrinsic evaluation measures, supporting their claims.

3 Bias Evaluation Measures

Several measures have already been proposed in prior work for evaluating the social biases encoded in MLMs. These measures can be broadly categorised into two groups depending on whether they

evaluate social biases in MLMs with respect to a particular downstream task or not. We refer to measures that evaluate social biases in MLMs on their own right, independently of any downstream tasks as *intrinsic MLM bias scores* and describe three such measures. In contrast, we refer to the measures that evaluate social biases in MLMs when they are applied to solve a specific downstream task such as Natural Language Inference (NLI), Semantic Textual Similarity (STS) or predicting occupations from biographies as *extrinsic MLM bias scores* and describe those measures.

To describe the intrinsic measures, let us consider a test sentence $S = w_1, w_2, \ldots, w_{|S|}$, containing length $|S|$ sequence of tokens w_i, where part of S is modified to create a stereotypical (or lack of thereof) example for a particular social bias. For example, let's consider the sentence-pair "***John completed his PhD in machine learning***" vs. "***Mary completed her PhD in machine learning***". Here, {*John*, *his*} are the modified tokens for the first sentence, whereas for the second sentence they are {*Mary*, *her*}. The unmodified tokens between the two sentences are {*completed*, *PhD*, *in*, *machine*, *learning*}.

Let us denote the list of modified tokens in a given sentence S by M and the list of unmodified tokens by U, so that $S = M \cup U$ is the list of all tokens in S. [1] Given an MLM with pretrained parameters θ to evaluate for social biases, let us denote the probability $P_{\text{MLM}}(w_i|S_{\setminus w_i}; \theta)$ that the MLM assigns to a token w_i conditioned on the remainder of the tokens, $S_{\setminus w_i}$. Salazar et al. (2020) demonstrated that $\text{PLL}(S)$, the pseudo-log-likelihood (PLL) score of sentence S given by (1), can be used to evaluate the preference expressed by an MLM for S, similarly to using log-probabilities for evaluating the naturalness of sentences using conventional language models.

$$\text{PLL}(S) \coloneqq \sum_{i=1}^{|S|} \log P_{\text{MLM}}(w_i|S_{\setminus w_i}; \theta) \quad (1)$$

PLL scores for MLMs can be computed directly and are more uniform across sentence lengths (no left-to-right bias), allowing us to recognize natural sentences in a language (Wang and Cho, 2019). As we will see later, PLL can be used to define bias evaluation scores for MLMs in a variety of ways.

---

[1] Note that to account for repeated occurrences of a word in a sentence, we examine lists rather than sets.

Intrinsic: StereoSet Score The probability of generating the modified tokens given the unmodified tokens in S was expressed as $P(M|U; \theta)$ by Nadeem et al. (2020). We refer to this as StereoSet Score (**SSS**) and it is given by (2).

$$\text{SSS}(S) \coloneqq \frac{1}{|M|} \sum_{w \in M} \log P_{\text{MLM}}(w|U; \theta) \quad (2)$$

$|M|$ denotes the length of M. SSS presents a challenge because, when comparing $P(M|U; \theta)$ for modified words like *John*, we might see high probabilities simply because these words occur frequently in the data used to train the MLM and not because the MLM has picked up a social bias.

Intrinsic: CrowS-Pairs Score The CrowS-Pairs Score (**CPS**) is a scoring formula provided by (3) that was defined as $P(U|M; \theta)$ by Nangia et al. (2020) to address the frequency-bias in SSS.

$$\text{CPS}(S) \coloneqq \sum_{w \in U} \log P_{\text{MLM}}(w|U_{\setminus w}, M; \theta) \quad (3)$$

Normalization is not performed here because the length of unmodified tokens is the same. However, by masking and predicting one token w at a time from U, we are effectively changing the context $(U_{\setminus w}, M)$ used as input by the MLM. This has two disadvantages. First, by removing w from the sentence, the MLM loses information that it can use to predict w. As a result, the prediction accuracy of w may decrease, rendering bias evaluations untrustworthy. Second, the remaining tokens $\{U_{\setminus w}, M\}$ may still be biased even if we remove one token w at a time from U. Moreover, the context on which we condition the probabilities continuously varies across predictions.

Intrinsic: All Unmasked Likelihood with Attention weights Kaneko and Bollegala (2022) proposed All Unmasked Likelihood with Attention weights (**AULA**) to overcome the above-mentioned disfluencies in previously proposed MLM bias evaluation measures. To begin, rather than masking out tokens from S, AULA provides the entire sentence to the MLM. Second, all tokens in S that appear in between the start and end of sentences are predicted by AULA. Furthermore, AULA computes the likelihood using attention weights to evaluate social biases based on the relative importance of

words in a sentence, as given by (4).

$$\text{AULA}(S) \coloneqq \frac{1}{|S|} \sum_{i=1}^{|S|} \alpha_i \log P_{\text{MLM}}(w_i|S;\theta) \tag{4}$$

Here, α_i represents the average of all multi-head attentions associated with w_i.

Given a score function $f \in \{\text{SSS}, \text{CPS}, \text{AULA}\}$, we use the percentage of stereotypical (S^{st}) test sentences preferred by the MLM over anti-stereotypical (S^{at}) ones to define the corresponding bias evaluation measure (**bias score**) as follows:

$$\left(\frac{100}{N} \sum_{(S^{\text{st}}, S^{\text{at}})} \mathbb{I}(f(S^{\text{st}}) > f(S^{\text{at}})) \right) - 50 \tag{5}$$

Here, the indicator function $\mathbb{I}$ returns 1 if its argument is True and 0 otherwise, and N is the total number of test instances. Values close to 0 indicate that the MLM under consideration is neither stereotypically nor anti-stereotypically biased; thus, it can be considered unbiased. Values less than 0 indicate a bias toward the anti-stereotypical group, while values greater than 0 indicate a bias toward the stereotypical group.

Extrinsic: BiasBios To evaluate the gender bias in occupation classification, De-Arteaga et al. (2019) proposed a dataset where they collected biographies written in English from Common Crawl. For this purpose, they first filter lines that begin with a name-like-pattern (i.e. a sequence of two capitalised words) followed by the string "is a(n) (xxx) *title*", where *title* is an occupation from the BLS Standard Occupation Classification system.[2] They used twenty-eight frequent occupations in Common Crawl to collect 397,340 biographies. The task is to predict the people's occupations, taken from the first sentence of their biographies, given the remainder of their biographies. For example, given the hypothetical biography of a nurse: *She graduated from Lehigh University, with honours in 1998. Nancy has years of experience in weight loss surgery, patient support, education, and diabetes.*, the goal is to predict the occupation.

An MLM under evaluation is used to create an embedding for a biography excluding the title sentence, and a softmax classifier is trained using those

biography embeddings to predict the occupation of the person described in each biography. If the MLM is unbiased with respect to gender, we would expect it to have approximately similar true positive rates (TPRs) averaged over all twenty-eight occupations for both male and female gendered persons. The TPR gap between the TPRs for classifying occupations in female and male biographies is used as an extrinsic MLM bias score. An unbiased MLM is expected to predict occupations with similar TPRs for both genders.

Extrinsic: STS-bias Semantic Textual Similarity (STS) is a task which evaluates the degree of semantic similarity between two sentences. Webster et al. (2020) proposed STS-bias to evaluate gender bias by adapting STS formulation. Specifically, to create the STS-bias dataset, they collect the 276 sentences from the STS-B test set, which start with *A man* or *A woman*, and discard the sentences with multiple gendered words.[3] For each template, they formed two sentence pairs per profession from Rudinger et al. (2018), one representing the male gender and the other for the female gender. For example, given a pair "*A man is walking*" and "*A nurse is walking*" from STS-B dataset, they create a new pair "*A woman is walking*" and "*A nurse is walking*". If not relying on gendered correlations, a model should assign equal estimates of similarity to the two pairs. Therefore, following Rudinger et al. (2017), as a score we use the difference between similarities of such two pairs to measure how similar model predictions actually are regarding gender. They compute the Pearson correlation between the scores per profession and the representation in that profession listed in the US Bureau of Labor Statistics (Caliskan et al., 2017). STS-bias does not consider the directionality of the gender bias. Therefore, we report absolute values of all bias scores.

Extrinsic: NLI-bias Dev et al. (2019) used natural language inference (NLI) for evaluating gender bias. This dataset contains sentence triples where a premise must be neutral in entailment w.r.t. two hypotheses. For example, "*The accountant ate a bagel*" should neither entail nor contradict the sentences "*The man ate a bagel*" and "*The woman ate a bagel*". Given a set containing $\mathcal{D}$ test instances, let the entailment predictor's probabilities

---

[2]https://www.bls.gov/soc

[3]Gender word list: https://github.com/uclanlp/gn_glove/tree/master/wordlist

for the d-th instance for entail, neutral and contradiction labels be respectively e_d, n_d and c_d. Then, they proposed the measure Fraction Neutral (FN):
$$\text{FN} = \frac{1}{|\mathcal{D}|} \sum_{d=1}^{\mathcal{D}} \mathbf{1}[\text{neutral} = \max(e_d, n_d, c_d)].$$
For an ideal (bias-free) embedding, FN measure would be 1. They compute the FN for each female and male sentences and use the difference between them as the bias score in the NLI-bias.

4 Experiments and Findings

4.1 Datasets

We used Crows-Pairs dataset (Nangia et al., 2020) as CPS and AULA evaluation data, and used StereoSet dataset (Nadeem et al., 2021) as SSS evaluation data. In BiasBios[4], we use the average performance gap in the male and female groups aggregated across all occupations to measure the bias, following Zhao et al. (2020a). We crawled BiasBios data using publicly available code from the original authors, and used 80% of the collected data as a training data, 10% as development data, and 10% as evaluation data. Because the STS-bias dataset is not publicly available, we created the dataset following to the procedure proposed by Webster et al. (2020). We used STS-B as training and development data (Cer et al., 2017)[5] for STS-bias, following Webster et al. (2020). We used the Multi-Genre Natural Language Inference (MNLI) as training data and development data (Williams et al., 2018)[6] for NLI-bias [7], following Kaneko and Bollegala (2021a).

4.2 Debiasing methods

We used three previously proposed methods that are widely used for debiasing MLMs such as All-layer Token-level debiasing (**AT**; Kaneko and Bollegala, 2021a), Counterfactual Data Augmentation debiasing (**CDA**; Webster et al., 2020), and dropout debiasing (**DO**; Webster et al., 2020). These methods represent conceptually different approaches for debiasing and can be applied to any pretrained MLM model/algorithm.

AT debiasing: Kaneko and Bollegala (2021a) proposed a method for debiasing MLMs via fine-tuning by simultaneously (a) preserving the semantic information in the pretrained MLM, and (b) re-

moving discriminative gender-related biases via an orthogonal projection in the hidden layers by operating at token- or sentence-levels, enabling debiasing on different layers and levels in the pre-trained contextualised embedding model. It is independent of model architectures and pre-training methods, thus can be adapted to a wide range of MLMs. They showed that applying token-level debiasing for all tokens and across all layers of an MLM produces the best performance. We perform an additional phase of fine-tuning using AT debiasing.

CDA debiasing: CDA is a debiasing strategy to mitigate social bias based on data (Webster et al., 2020). CDA involves re-balancing a corpus by swapping bias attribute words in a dataset. For example, the sentence "*the doctor treated his patient*" could be changed to "*the doctor treated her patient*". The re-balanced corpus is then often used for further training to debias a model. We experiment with debiasing pre-trained MLMs by performing an additional phase of fine-tuning on counterfactually augmented sentences.

DO debiasing Webster et al. (2020) proposed dropout regularization for a bias mitigation. They increased the dropout parameters for MLM's attention weights and hidden activations and further pre-trained them. They showed that increased dropout regularization reduces gender bias within these MLMs. They hypothesize that dropout's interruption of the attention mechanisms within MLMs prevents MLMs from learning undesirable associations between words. Similar to AT and CDA, we perform an additional phase of fine-tuning using increased dropout regularization.

We used news-commentary-v15[8] corpus as additional training data for debiasing. We used publicly available code[9] and default hyperparameters for AT debiasing. We used Zhao et al. (2018)'s word list to replace a feminine word to the corresponding masculine word and vice versa in sentences in the training data for CDA debiasing. We set the dropout ratio to 0.15 for attention probabilities and 0.2 for all fully connected layers in DO debiasing. Hyperparameters of CDA and DO debiasing are set to their default values in the Huggingface Transformers library (Wolf et al., 2020).

---

[4]https://github.com/microsoft/biosbias
[5]github.com/facebookresearch/SentEval
[6]cims.nyu.edu/~sbowman/multinli/
[7]https://github.com/sunipa/On-Measuring-and-Mitigating-Biased-Inferences-of-Word-Embeddings

[8]https://data.statmt.org/news-commentary/v15/
[9]https://github.com/kanekomasahiro/context-debias

4.3 Models and Hyperparameters

In our experiments, we used the following seven MLMs: bert-base-uncased (**bert-bu**)[10], bert-base-cased (**bert-bc**),[11] bert-large-uncased (**bert-lu**),[12] bert-large-cased (**bert-lc**),[13] roberta-base (**roberta-b**),[14] roberta-large (**roberta-l**),[15] albert-base (**albert-b**).[16]

To fine-tune an MLM for a downstream task, we first create an embedding for a training instance using the given MLM. Specifically, we use the [CLS] token embedding produced by the MLM to represent an instance. For STS and NLI tasks where an instance consists of a pair of sentences, we concatenate the two sentences in the pair and use it as the input to the MLM. Finally, a classification head (i.e. a softmax layer) is used to predict the label for a training instance, and cross entropy error is backpropagated to update the MLM parameters.

For downstream tasks, the best performance checkpoint on development data is selected from $\{16, 32, 64\}$ for batch size, {1e-5, 3e-5, 5e-5} for learning rate, $\{1, 3, 5\}$ for epoch number with greedy search. Maximum sentence length is set to 128 tokens and we used four Tesla V100 GPUs in our experiments. All other hyperparameters, are set to their default values in the huggingface library.

4.4 Debiasing vs. Bias in Downstream Tasks

We examine whether the debiasing methods proposed for MLMs can mitigate the social biases when those debiased MLM are used in downstream tasks. Specifically, given an MLM, M, we use each debiasing method described in §4.2 independently on M to create a debiased version, M_{deb}. Let, $f(M)$, denote the bias score of M measured using a bias evaluation measure, f, described in §3. Figure 1 shows the reduction of bias $f(M_{\mathrm{deb}}) - f(M)$ due to debiasing for multiple MLMs and debiasing methods for different f. For all debiasing methods, we see a reduction in biases according to intrinsic bias evaluation measures SSS, CPS and AULA. However, among the extrinsic bias evaluation measure, we see a reduction according to STS, but an *increase* in biases according to BiasBios and NLI. For example, AT for **bert-bu** mitigates bias

---

[10]huggingface.co/bert-base-uncased
[11]huggingface.co/bert-base-cased
[12]huggingface.co/bert-large-uncased
[13]huggingface.co/bert-large-cased
[14]huggingface.co/roberta-base
[15]huggingface.co/roberta-large
[16]huggingface.co/albert-base-v2

	Feminine	Masculine	ILSP
BiasBios	20,720	35,412	-0.53
STS-B	1,304	2,126	-0.48
MNLI	55,588	139,575	-0.91

Table 1: Frequency of feminine and masculine words in downstream task data. ILSP represents the degree of bias between feminine and masculine words.

in all intrinsic measures, but increases biases in all extrinsic measures. This shows that debiasing methods do *not* always reduces extrinsic biases, despite appearing to be debiased with respect to intrinsic measures. Note that the training dataset sizes for BioBias, STS-B and MLNI are respectively 251K, 5.7K, and 392.7K. Therefore, STS-B has fewer training data instances than for the other two tasks, and the number of fine-tuning iterations required is also smaller. For these reasons, we believe that debiased MLMs are less influenced when fine-tuned on STS-B.

5 Why is Debiasing less Effective for Downstream Tasks?

To further investigate the discrepancies between intrinsic and extrinsic bias evaluation measures observed in §4.4, recall that a debiased MLM can re-learn social biases in two different sources via fine-tuning: *instance-related* biases and *label-related* biases. A training data point of a downstream task consists of an *instance* (i.e. sentence or text) and a *label* (i.e. entailment labels, similarity ratings etc.). Instances alone could express social biases such as in the sentence "*All Muslims are terrorists*". By fine-tuning a debiased MLM on such instances using the masking training objective, it can learn racially discriminatory associations. We call this the instance-related bias in fine-tuning.

Second, even if an instance is unbiased, coupled with its label, the data point could still express a social bias. For example, the sentence pair ("*A nurse is walking*", "*A woman is walking*") represents an instance, which does not represent a gender bias. However, if we are told that the first sentence entails the second (i.e. an entailment label assigned to this instance), it will represent a gender bias. An NLP model can easily learn such label-related biases in an end-to-end training setting by propagating the prediction losses all the way back to the MLM layers.

To examine the instance-related gender biases in

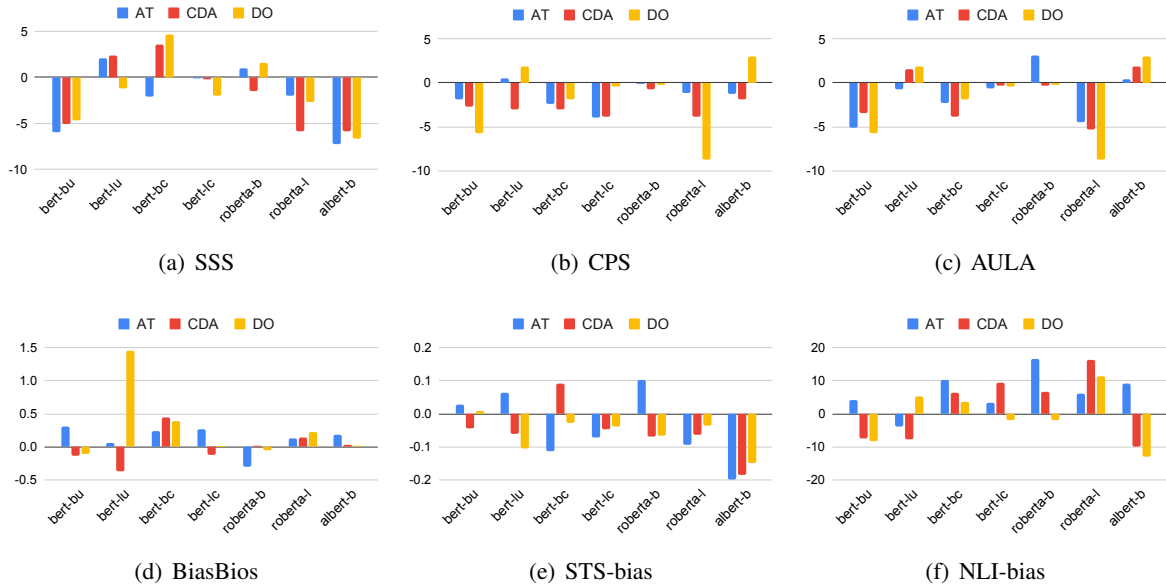| (a) SSS	(b) CPS	(c) AULA
(d) BiasBios	(e) STS-bias	(f) NLI-bias

Figure 1: Differences between the bias scores of original vs. debiased MLMs. Negative values indicate that the debiased MLM has a lower bias than its original (non-debiased) version.

downstream training data, we compute ILPS (Kurita et al., 2019) given by $\log(P_f/P_m)$, where P_f and P_m are the probabilities of respectively female and male genders in training instances. Prior work has shown that the frequency imbalance of feminine and masculine words in the training data is related to the gender bias learnt by MLMs (Kaneko and Bollegala, 2022; Kaneko et al., 2022b). Using the gender word lists created by Zhao et al. (2018), we estimate ILPS as shown in Table 1.

From Table 1, we see that all three datasets are highly biased towards the male gender, indicated by the negative ILSP scores. If we arrange the three datasets in the ascending order of their gender biases we get: STS-B < BiasBios < MLNI. Interestingly, this is exactly the same order we would get if we had ordered these datasets considering the number of MLMs that have their biases increased due to fine-tuning in Figure 1 (e, d, f), respectively.

5.1 Re-learning Biases via Fine-Tuning

To further study how biases are learnt during the fine-tuning process, we track the extrinsic bias score of an MLM (and its debiased variants) over the number of fine-tuning training iterations on a particular downstream task. Here, we use **bert-bu** and **roberta-l**, which had low intrinsic biases after debiasing according to all intrinsic bias measures in Figure 1. We save model checkpoints for every 500 fine-tuning iterations up to 5500 iterations and evaluate their extrinsic bias scores.

Figure 2 shows that the extrinsic bias scores of non-debiased (original) MLMs (**roberta-l** and **bert-bu**) and their debiased variants using AT, CDA, and DO. On all downstream tasks, extrinsic bias scores increase during fine-tuning for both non-debiased and debiased MLMs. This shows that biases that were mitigated by debiasing are re-learnt during fine-tuning on downstream training data. Even STS-bias, which reported almost all MLMs to be void of biases after debiasing (Figure 1), shows increasing bias scores during fine-tuning. Overall, we see that debiasing effect gradually decreases as fine-tuning progresses due to the biases in downstream task data. In contrast, Jin et al. (2021) argued that debiased MLMs are unaffected by fine-tuning. However, their evaluation was limited to only one MLM and two datasets both related to gender stereotyping (Jin et al., 2021), whereas we consider seven MLMs, three debiasing methods and six bias evaluation measures. This shows that careful evaluation using various evaluation datasets and MLMs is necessary when verifying the effectiveness of debiasing methods.

Furthermore, according to the intrinsic evaluations in Figure 1, the biases were reduced in both **roberta-l** and **bert-bu** by all debiasing methods. However, there are cases where the bias scores of debiased MLMs are higher than that of non-debiased MLMs even in the early stages of fine-tuning, except for **roberta-l** in the STS-bias in Fig-
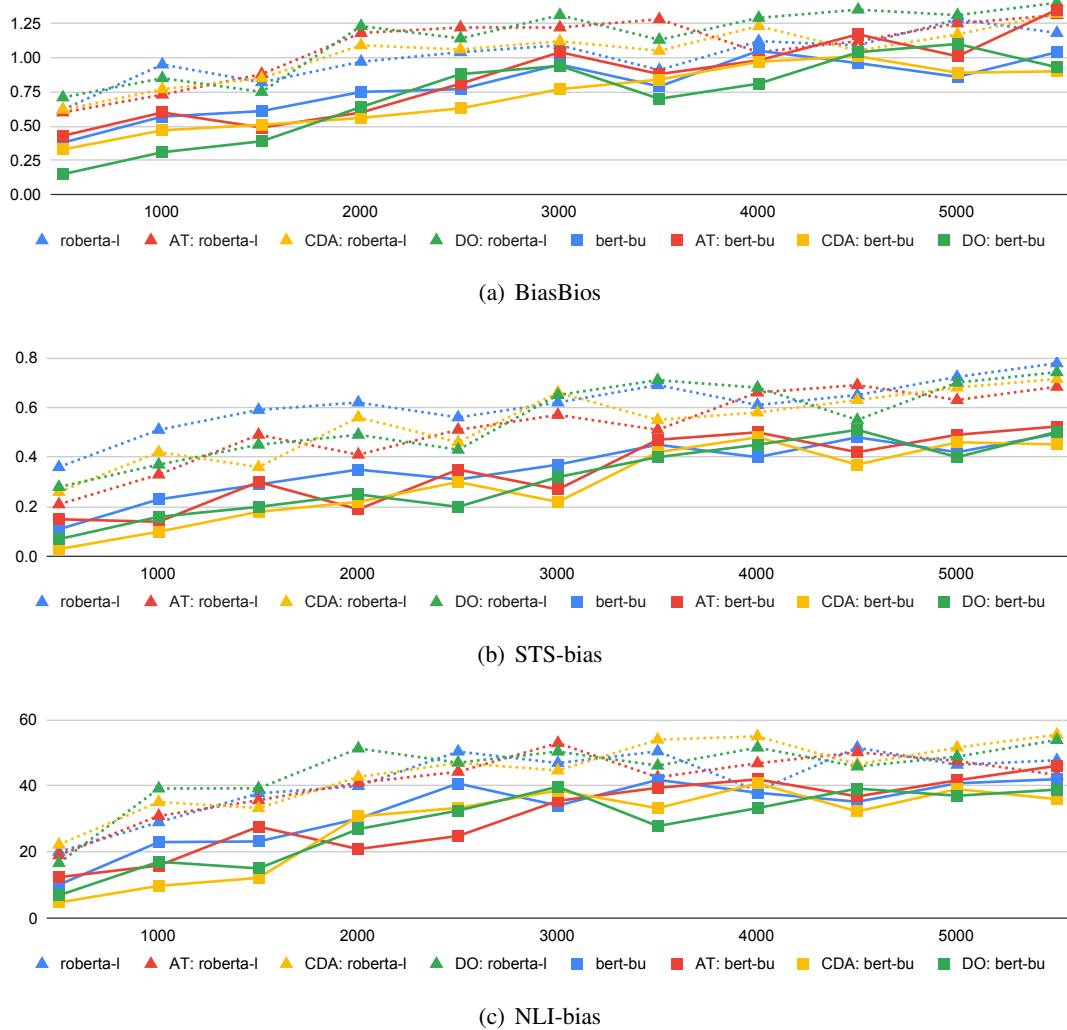
(a) BiasBios

(b) STS-bias

(c) NLI-bias

Figure 2: Extrinsic bias scores measured for MLMs during fine-tuning on downstream instances.

ure 2. Therefore, MLMs judged to be debiased according to intrinsic evaluations alone might not necessarily remain debiased after fine-tuning on downstream task instances.

5.2 Correlations between Intrinsic vs. Extrinsic Bias Evaluation Measures

Given that intrinsic bias evaluation measures are used to determine whether an MLM should be used in a downstream application, it is important to examine what level of correlations exist between intrinsic and extrinsic bias evaluation measures (Blodgett et al., 2021). For this purpose, using 28 MLMs (original 7 MLMs and their debiased versions using three methods, $7 \times 3 + 7 = 28$) we compute the Kendall's τ rank correlation coefficients between intrinsic and extrinsic bias measures. Specifically, in Table 2 we show τ values between the intrinsic bias measure (SSS, CPS, AULA) and

extrinsic measures (BiasBios, STS-bias, NLI-bias). In each cell in Table 2, the values to the left of the slashes (/) are computed using the MLMs prior to fine-tuning on downstream data, whereas the values to the right of the slashes are computed using the MLMs that have been fine-tuned on instances of downstream data.[17] Statistically significant ($p < 0.05$) τ values according to the Fisher's exact test are denoted by a †. Cases where the correlation has improved after fine-tuning are highlighted in **Bold**.

From Table 2, we see that prior to fine-tuning, only two τ values were significant, reconfirming the weak correlations observed in prior work (Delobelle et al., 2021; Goldfarb-Tarrant et al., 2021; Cao et al., 2022). Given that intrinsic bias evaluation measures are agnostic to the biases in downstream

---

[17]We use the fine-tuned models after 5500 iterations in Figure 2.

	SSS	CPS	AULA
BiasBios	-0.02 / **0.24**	-0.17 / -0.17	-0.20 / **0.11**
STS-bias	0.53† / 0.38†	-0.08 / **0.12**	0.24 / 0.13
NLI-bias	0.22 / 0.21	0.16 / 0.05	0.30† /0.17

Table 2: Kendall τ correlation coefficients between intrinsic extrinsic bias scores. Values to the left and right of '/' are computed using MLMs respectively prior and post fine-tuning. Statistically significant ($p < 0.05$) τ values are denoted by a $\dagger$, whereas improved correlations after fine-tuning are highlighted in **Bold**.

task-specific training data, one might have expected to see the correlations between intrinsic and extrinsic measures to improve *after* the MLMs have been fine-tuned on downstream data, thereby capturing some of the biases in downstream data within the MLMs. Surprisingly, we see that this is clearly not the case here and only in three out of nine combinations in Table 2 we see correlations improving after fine-tuning. In particular, for NLI-bias we see that correlations dropping after fine-tuning against all intrinsic measures compared to their values prior to fine-tuning. On the other hand, for BiasBios we see that correlations improving after fine-tuning against SSS and AULA by large margins, while remaining unaffected against CPS. The only significant correlation after fine-tuning is between STS-bias and SSS (i.e. 0.38). However, this too has decreased from its value prior to fine-tuning (i.e. 0.53).

Overall, our results indicate weak and inconsistent correlations between intrinsic and extrinsic bias evaluation measures irrespective of whether the MLMs have been fine-tuned on downstream data or otherwise. Therefore, we recommend that intrinsic evaluation scores alone must not be used to determine whether an MLM is sufficiently unbiased to be deployed to NLP systems that are used by billions of users world-wide.

6 Conclusion

We investigated social biases in 28 MLMs and found that there exist weak correlation between intrinsic and extrinsic bias evaluation measures. Moreover, debiased MLMs re-learn biases during task-specific fine-tuning due to instance and label related biases in downstream task data. Our findings highlight limitations in existing bias evaluation measures and raise serious concerns regarding their implications.

7 Ethical Considerations

Our goal in this paper was to evaluate the correlation between previously proposed and widely-used intrinsic and extrinsic bias evaluation measures for evaluating various types of social biases in pretrained MLMs. However, we did not manually annotate novel social bias datasets or proposed novel bias evaluation measures nor debiasing methods. Therefore, we do not see any ethical issues arising due to data annotation, or via proposals of novel evaluation metrics or debiasing methods. However, we would like to point out that although we conducted an extensive study using 7 MLMs, 3 debiasing methods, 3 intrinsic evaluation measures and 3 extrinsic evaluation measures, this is still a limited subset of all possible scenarios. For example, the gender biases we considered in this paper cover only binary gender (Dev et al., 2021). However, non-binary genders are severely underrepresented in textual data used to train MLMs. Moreover, non-binary genders are often associated with derogatory adjectives. Evaluating the correlation between intrinsic and extrinsic non-binary gender is an important next step.

All MLMs used in this study are trained for English. However, as reported in much prior work social biases are language independent and omnipresent in MLMs trained for many languages (Kaneko et al., 2022b; Lewis and Lupyan, 2020; Liang et al., 2020b; Bartl et al., 2020; Zhao et al., 2020b). We plan to extend this study to cover non-English MLMs in the future.

Kaneko et al. (2022a) show that combining multiple debiasing methods for word embeddings (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019, 2021b, 2020; Ravfogel et al., 2020) can complement each other's strengths and weaknesses to obtain more effective debiased embeddings. Therefore, combining methods to mitigate both intrinsic and extrinsic bias may allow for more effective debiasing.

Finally, biases are not limited to word representations but also appear in sense representations (Zhou et al., 2022). On the other hand, our correlation analysis did not include any sense embedding models.

References

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and

mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *NAACL-HLT*, pages 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

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

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. In *Proc. of ACL*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 120–128, New York, NY, USA. Association for Computing Machinery.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447*.

Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2019. On Measuring and Mitigating Biased Inferences of Word Embeddings.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. 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. In *NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 862–872, New York, NY, USA. Association for Computing Machinery.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2019. Addressing age-related bias in sentiment analysis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6146–6150. International Joint Conferences on Artificial Intelligence Organization.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.

Wei Guo and Aylin Caliskan. 2021. *Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases*, pages 122–133. Association for Computing Machinery, New York, NY, USA.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5266–5274, Hong Kong, China. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proc. of ACL*, pages 591–598.

Conversational AI Jigsaw. 2019. Unintended bias in toxicity classification. Kaggle.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings*

of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3770–3783, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *ACL*, pages 1641–1650.

Masahiro Kaneko and Danushka Bollegala. 2020. Autoencoding improves pre-trained word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1699–1713, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021a. Debiasing pre-trained contextualised embeddings. In *Proc. of 16th conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Masahiro Kaneko and Danushka Bollegala. 2021b. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask – evaluating social biases in masked language models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, page (to appear), Vancouver, BC, Canada.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022a. Gender bias in meta-embeddings. *arXiv preprint arXiv:2205.09867*.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022b. Gender bias in masked language models for multiple languages. In *Proc. of NAACL-HLT*.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv*, abs/1909.11942.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10):1021–1028.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020a. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020b. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proc. of 35th AAAI Conference on Artificial Intelligence (AAAI)*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. 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 *Proc. of NAACL-HLT*.

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

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020a. Gender bias in multilingual embeddings and cross-lingual transfer.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020b. Gender bias in multilingual embeddings and cross-lingual transfer. In *COLING*, pages 2896–2907. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *EMNLP*, pages 4847–4853.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense embeddings are also biased–evaluating social biases in static and contextualised sense embeddings. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proc. of ACL*.