An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models

Anonymous ACL submission

Abstract

Recent work has shown pre-trained language models capture social biases from the text corpora they are trained on. This has attracted attention to developing techniques that mitigate such biases. In this work, we perform an empirical survey of five recently proposed bias mitigation techniques: Counterfactual Data Augmentation (CDA), Dropout, Iterative Nullspace Projection, Self-Debias, and SentenceDebias. We quantify the effectiveness of each technique using three intrinsic bias benchmarks while also measuring the impact of these techniques on a model's language modeling ability, as well as its performance on downstream NLU tasks. We experimentally find that: (1) Self-Debias is the strongest debiasing technique, obtaining improved scores on all bias benchmarks; (2) Current debiasing techniques perform less consistently when mitigating non-gender biases; And (3) improvements on bias benchmarks such as StereoSet and CrowS-Pairs by using debiasing strategies are often accompanied by a decrease in language modeling ability, making it difficult to determine whether the bias mitigation was effective.[1]

1 Introduction

Large pre-trained language models have proven effective across a variety of tasks in natural language processing, often obtaining state of the art performance (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020). These models are typically trained on large amounts of text, originating from unmoderated sources, such as the internet. While the performance of these pre-trained models is remarkable, recent work has shown that they capture social biases from the data they are trained on (May et al. 2019; Kurita et al. 2019; Webster et al. 2020; Nangia et al. 2020; Nadeem et al. 2021, *inter alia*). Because of these findings, an increasing amount of research has focused on developing techniques to mitigate these biases (Liang et al., 2020; Ravfogel et al., 2020; Webster et al., 2020; Kaneko and Bollegala, 2021; Schick et al., 2021; Lauscher et al., 2021). However, the proposed techniques are often not investigated thoroughly. For instance, much work focuses *only* on mitigating gender bias despite pre-trained language models being plagued by other social biases (e.g., *racial* or *religious* bias). Additionally, the impact that debiasing has on both downstream task performance, as well as language modeling ability, is often not well explored.

In this paper, we perform an empirical survey of the effectiveness of five recently proposed debiasing techniques for pre-trained language models:[2] **Counterfactual Data Augmentation** (CDA; Zmigrod et al. 2019; Webster et al. 2020), **Dropout** (Webster et al., 2020), **Iterative Nullspace Projection** (INLP; Ravfogel et al. 2020), **Self-Debias** (Schick et al., 2021) and **SentenceDebias** (Liang et al., 2020). Following the taxonomy described by Blodgett et al. (2020), our work studies the effectiveness of these techniques in mitigating *representational biases* from pre-trained language models. More specifically, we investigate mitigating *gender*, *racial*, and *religious* biases in three masked language models (BERT, ALBERT, and RoBERTa) and an autoregressive language model (GPT-2). We also explore how debiasing impacts a model's language modeling ability, as well as a model's performance on downstream natural language understanding (NLU) tasks.

Concretely, our paper aims to answer the following research questions:

Q1 Which technique is most effective in mitigating bias?

[1] Our code is included with our submission and will be made publicly available.

[2] We select these techniques based upon popularity, ease of implementation, and ease of adaptation to non-gender biases.

Q2 Do these techniques worsen a model's language modeling ability?

Q3 Do these techniques worsen a model's ability to perform downstream NLU tasks?

To answer Q1 (§4), we evaluate debiased models against three intrinsic bias benchmarks: the Sentence Encoder Association Test (SEAT; May et al. 2019), StereoSet (Nadeem et al., 2021), and Crowdsourced Stereotype Pairs (CrowS-Pairs; Nangia et al. 2020). Generally, we found Self-Debias to be the strongest bias mitigation technique. To answer Q2 (§5) and Q3 (§6), we evaluate debiased models against WikiText-2 (Merity et al., 2016) and the General Language Understanding Evaluation (GLUE; Wang and Cho 2019) benchmark. We found debiasing tends to worsen a model's language modeling ability. However, our results suggest that debiasing has little impact on a model's ability to perform downstream NLU tasks.

2 Techniques for Measuring Bias

We begin by describing the three intrinsic bias benchmarks we use to evaluate our debiasing techniques. We select these benchmarks as they can be used to measure not only gender bias, but also *racial* and *religious* bias in language models.

Sentence Encoder Association Test (SEAT). We use SEAT (May et al., 2019) as our first intrinsic bias benchmark. SEAT is an extension of the Word Embedding Association Test (WEAT; Caliskan et al. 2017) to sentence-level representations. Below, we first describe WEAT.

WEAT makes use of four sets of words: two sets of bias *attribute* words and two sets of *target* words. The attribute word sets characterize a type of bias. For example, the attribute word sets {*man*, *he*, *him*, ...} and {*woman*, *she*, *her*, ...} could be used for gender bias. The target word sets characterize particular concepts. For example, the target word sets {*family*, *child*, *parent*, ...} and {*work*, *office*, *profession*, ...} could be used to characterize the concepts of *family* and *career*, respectively. WEAT evaluates whether the representations for words from one particular attribute word set tend to be more closely associated with the representations for words from one particular target word set. For instance, if the representations for the *female* attribute words listed above tended to be more closely associated with the representations for the *family* target words, this may be indicative of bias within the word representations.

Formally, let A and B denote the sets of attribute words and let X and Y denote the sets of target words. The SEAT test statistic is

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

where for a particular word w, $s(w, A, B)$ is defined as the difference between w's mean cosine similarity with the words from A and w's mean cosine similarity with the words from B

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(w, a) - \frac{1}{|B|} \sum_{b \in B} \cos(w, b).$$

They report an effect size given by

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

where μ denotes the mean and σ denotes the standard deviation. Here, an effect size closer to zero is indicative of a smaller degree of bias in a model's representations.

To create a sentence-level version of WEAT (referred to as SEAT), May et al. (2019) substitute the attribute words and target words from WEAT into synthetic sentence templates (e.g., "*this is a [WORD]*") to create a collection of sentences. Now, given sets of sentences containing *attribute* and *target* words, the WEAT test statistic can be computed using sentence-level representations obtained from a pre-trained language model.[3]

We refer readers to Appendix A for a list of the SEAT tests we use to measure each type of bias in our work. We report the effect size for each SEAT test we evaluate.

StereoSet. As our second intrinsic bias benchmark, we use StereoSet (Nadeem et al., 2021), a crowdsourced dataset for measuring four different types of stereotypical bias in language models. Each StereoSet example consists of a context sentence, for example "*our housekeeper is [MASK]*", and a set of three candidate associations (completions) for that sentence—one being stereotypical, another being anti-stereotypical, and a third being unrelated.[4] Using the example

[3] We use a permutation on the SEAT test statistic to compute the significance of association between the attribute word sets and the target word sets. We refer readers to the original work of Caliskan et al. (2017) for a complete description of this test.

[4] We consider only the *intrasentence* task from StereoSet. Thus, henceforth, when we refer to a StereoSet example, we are referring to a StereoSet *intrasentence* example.

above, a stereotypical association might be "*our housekeeper is **Mexican***", an anti-stereotypical association might be "*our housekeeper is **American***", and an unrelated association might be "*our housekeeper is **computer***". To quantify how biased a language model is, we score the stereotypical association and the anti-stereotypical association for each example under a model. We then compute the percentage of examples for which a model prefers the stereotypical association as opposed to the anti-stereotypical association. We define this percentage as the **stereotype score** of a model.

StereoSet also provides a measure of a model's language modeling ability. For each example in the dataset, we also score the *unrelated* association. We then measure the percentage of examples for which a model prefers a meaningful association (either the stereotypical association or the anti-stereotypical association) as opposed to the *unrelated* association. We define this percentage as the **language modeling score** of a model.

We evaluate our debiased models against the StereoSet test set. We evaluate debiased models for each domain against their respective StereoSet test set split (e.g., *gender* debiased models are evaluated against the gender bias examples).

Crowdsourced Stereotype Pairs (CrowS-Pairs). We use CrowS-Pairs (Nangia et al., 2020) as our third intrinsic bias benchmark. CrowS-Pairs is a crowdsourced dataset that consists of pairs of minimally distant sentences—that is, sentences that differ only with respect to a small number of tokens. The first sentence in each pair reflects a stereotype about a historically disadvantaged group in the United States. For example, the sentence "*people who live in trailers are alcoholics*" reflects a possible socioeconomic stereotype. The second sentence in each pair then *violates* the stereotype introduced in the first sentence. For example, the sentence "*people who live in **mansions** are alcoholics*" violates, or in a sense, is the anti-stereotypical version of the first sentence.

We quantify how biased a language model is by measuring how frequently a model prefers the stereotypical sentence in each pair over the anti-stereotypical sentence. Nangia et al. (2020) originally proposed using pseudo-likelihood-based scoring (Salazar et al., 2020) for CrowS-Pairs, however, recent work has suggested that pseudo-likelihood-based scoring may be subject to model calibration issues (Desai and Durrett, 2020; Jiang et al., 2020). Thus, we score each pair of sentences using masked token probabilities in a similar fashion to StereoSet. For each pair of sentences, we score the stereotypical sentence by computing the masked token probability of the tokens unique to the stereotypical sentence. In the example above, we would compute the masked token probability of *trailers*. We score each anti-stereotypical sentence in a similar fashion. If multiple tokens are unique to a given sentence, we compute the *average* masked token probability by masking each differing token individually. We define the **stereotype score** of a model to be the percentage of examples for which a model assigns a higher masked token probability to the stereotypical sentence as opposed to the anti-stereotypical sentence.

3 Debiasing Techniques

Below, we describe the five debiasing techniques we evaluate in this work. We refer readers to Appendix C for additional experimental details on each debiasing technique.

CDA. CDA (Zmigrod et al., 2019; Dinan et al., 2020; Webster et al., 2020; Barikeri et al., 2021) is a data-based debiasing strategy that is often used to mitigate gender bias. Roughly, CDA involves *re-balancing* a corpus by *swapping* bias attribute words (e.g., *he*/*she*) in a dataset. For instance, to help mitigate gender bias, the sentence "*the doctor went to the room and **he** grabbed the syringe*" could be augmented to "*the doctor went to the room and **she** grabbed the syringe*". The re-balanced corpus is then often used for further training to debias a model. While CDA has been mainly used for gender debiasing, we also evaluate its effectiveness for other types of biases. For instance, we create CDA data for mitigating *religious* bias by swapping religious terms in a corpus, say *church* with *mosque*, to generate counterfactual examples.

We experiment with debiasing pre-trained language models by performing an additional phase of pre-training on counterfactually augmented sentences from English Wikipedia.[5]

DROPOUT. Webster et al. (2020) investigate using dropout regularization (Srivastava et al., 2014) as a bias mitigation technique. They investigate increasing the dropout parameters for BERT and ALBERT's attention weights and

[5]We provide the bias attribute words we make use of in our study in Appendix B.

hidden activations and performing an additional phase of pre-training. They find using increased dropout regularization reduces gender bias within these models. They hypothesize that dropout's interruption of the attention mechanisms within BERT and ALBERT help prevent them from learning undesirable associations between words. We extend this study to other types of biases. Similar to CDA, we perform an additional phase of pre-training on sentences from English Wikipedia using increased dropout regularization.

SELF-DEBIAS. Schick et al. (2021) propose a post-hoc debiasing technique that leverages a model's internal knowledge to discourage it from generating biased text.

Informally, Schick et al. (2021) propose using manually curated prompts to first *encourage* a model to generate toxic text. For instance, generation from an autoregressive model could be prompted with "*The following text discriminates against people because of their* ***gender***." Then, a *second* continuation that is non-discriminative can be generated from the model where the probabilities of tokens deemed likely under the first toxic generation can be scaled down.

Importantly, since Self-Debias is a post-hoc text generation debiasing procedure, it does not alter a model's internal representations or its parameters. Thus, Self-Debias cannot be used as a bias mitigation strategy for downstream NLU tasks (e.g., GLUE). Additionally, since SEAT measures bias in a model's representations and Self-Debias does not alter a model's internal representations, we cannot evaluate Self-Debias against SEAT.

SENTENCEDEBIAS. Liang et al. (2020) extend *Hard-Debias*, a word embedding debiasing technique proposed by Bolukbasi et al. (2016) to sentence representations. SentenceDebias is a projection-based debiasing technique that requires the estimation of a linear subspace for a particular type of bias. Sentence representations can be debiased by projecting onto the estimated bias subspace and subtracting the resulting projection from the original sentence representation.

Liang et al. (2020) use a three step procedure for computing a bias subspace. First, they *define* a list of bias attribute words (e.g., *he*/*she*). Second, they *contextualize* the bias attribute words into sentences. This is done by finding occurences of the bias attribute words in sentences within a text corpus. For each sentence found during this contextualization step, CDA is applied to generate a pair of sentences that differ only with respect to the bias attribute word. Finally, they *estimate* the bias subspace. For each of the sentences obtained during the contextualization step, a corresponding representation can be obtained from a pre-trained model. Principle Component Analysis (PCA; Abdi and Williams 2010) can then be used to estimate the principle directions of variation of the resulting set of representations. The first K principle components can be taken to define the bias subspace.

INLP. Ravfogel et al. (2020) propose Iterative Nullspace Projection (INLP), a projection-based debiasing technique similar to SentenceDebias. Roughly, INLP debiases a model's representations by training a linear classifier to *predict* the protected property you want to remove (e.g., gender) from the representations. Then, representations can be debiased by projecting them into the nullspace of the learnt classifier's weight matrix, effectively removing all of the information the classifier used to predict the protected attribute from the representation. This process can then be applied iteratively to debias the representation.

In our experiments, we create a classification dataset for INLP by finding occurrences of bias attribute words (e.g., *he*/*she*) in English Wikipedia. For example, for gender bias, we classify each sentence from English Wikipedia into one of three classes depending upon whether a sentence contains a *male* word, a *female* word, or *no* gendered words.

4 Which Technique is Most Effective in Mitigating Bias?

To investigate which technique is most effective in mitigating bias (**Q1**), we evaluate debiased BERT, ALBERT, RoBERTa, and GPT-2 models against SEAT, StereoSet, and CrowS-Pairs. We present BERT and GPT-2 results in the main paper and defer readers to Appendix E for results for the other models. We use the *base uncased* BERT model and the *small* GPT-2 model in our experiments.

SEAT Results. In Table 1, we report results for gender debiased BERT and GPT-2 models on SEAT.

For BERT, we find all of our debiased models obtain lower average absolute effect sizes than the baseline model—an encouraging result. In

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	0.535*	0.056	-0.925	0.352	0.303	0.129	↓0.237 0.383
+ DROPOUT	0.750*	0.189	-0.507	0.488*	0.348	0.202	↓0.206 0.414
+ INLP	0.551*	-0.160	-0.638	0.291	0.346	0.195	↓0.257 0.363
+ SENTENCEDEBIAS	0.350	-0.298	-0.623	0.464*	0.414	0.464*	↓0.185 0.435
GPT-2	0.138	0.003	-0.023	0.002	-0.224	-0.287	0.113
+ CDA	0.161	-0.034	0.898*	0.874*	0.516*	0.396	↑0.367 0.480
+ DROPOUT	0.167	-0.040	0.866*	0.873*	0.527*	0.384	↑0.363 0.476
+ INLP	0.300	0.365	-0.075	-0.137	-0.373	-0.384	↑0.160 0.273
+ SENTENCEDEBIAS	0.087	-0.072	-0.294	-0.064	0.318	-0.667	↑0.137 0.250

Table 1: **SEAT effect sizes for gender debiased BERT and GPT-2 models.** Effect sizes closer to 0 are indicative of less biased model representations. Statistically significant effect sizes at $p < 0.01$ are denoted by *. The final column reports the average absolute effect size across all 6 gender SEAT tests for each debiased model.

Model	Avg. Effect Size (↓)
Race	
BERT	0.620
+ CDA	↓0.322 0.298
+ DROPOUT	↓0.389 0.231
+ INLP	↑0.020 0.640
+ SENTENCEDEBIAS	↓0.008 0.612
GPT-2	0.448
+ CDA	↓0.309 0.139
+ DROPOUT	↓0.286 0.162
+ INLP	↓0.057 0.391
+ SENTENCEDEBIAS	↓0.031 0.417
Religion	
BERT	0.492
+ CDA	↓0.243 0.249
+ DROPOUT	↓0.269 0.223
+ INLP	↓0.031 0.461
+ SENTENCEDEBIAS	↓0.054 0.438
GPT-2	0.376
+ CDA	↓0.238 0.138
+ DROPOUT	↓0.242 0.134
+ INLP	↑0.018 0.394
+ SENTENCEDEBIAS	↑0.169 0.545

Table 2: **SEAT average absolute effect sizes for race and religion debiased BERT and GPT-2 models.** Average absolute effect sizes closer to 0 are indicative of less biased model representations.

particular, INLP performs best on average across all six SEAT tests. Interestingly, we note that CDA outperforms SentenceDebias. We found this result surprising as SentenceDebias takes a more aggressive approach to debiasing than CDA by attempting to remove *all* gender information from a model's representations.

For GPT-2, our results are much less encouraging. We find all of the debiased models obtain *higher* average absolute effect sizes than the baseline model. However, we note that SEAT fails to detect any statistically significant bias in the baseline model in any of the six SEAT tests to begin with. We argue, alongside others (Kurita et al., 2019; May et al., 2019), that SEAT's failure to detect bias in GPT-2 brings into question its reliability as a bias benchmark. For our gender debiased ALBERT and RoBERTa models, we observed similar trends in performance to BERT.

We also use SEAT to evaluate racial and religious bias in our models. In Table 2, we report average absolute effect sizes for race and religion debiased BERT and GPT-2 models. We find most of our race and religion debiased BERT and GPT-2 models obtain lower average absolute effect sizes than their respective baseline models. We observed a similar trend in the performance of our ALBERT and RoBERTa models.

StereoSet Results. In Table 3, we report StereoSet results for BERT and GPT-2.

For BERT, all of the gender debiased models obtain lower stereotype scores than the baseline model. However, the race and religion debiased models do not perform as consistently well. We note that for race, two of the five debiased models obtain lower stereotype scores than the baseline model and for religion, four of the five debiased models obtain lower stereotype scores than the baseline model. We observed similar trends to BERT in our ALBERT and RoBERTa results.

For GPT-2, the gender debiased models do not perform as consistently well. Notably, we observe that the CDA model obtains a higher stereotype score than the baseline model.

One encouraging trend in our results is the consistently strong performance of Self-Debias. Across all three bias domains, the Self-Debias BERT and GPT-2 models always obtain reduced stereotype scores. Similarily, four of the six Self-

Model	Stereotype Score (%)
Gender	
BERT	60.28
+ CDA	↓2.51 57.77
+ DROPOUT	↓0.99 59.29
+ INLP	↓0.49 59.79
+ SELF-DEBIAS	↓0.94 59.34
+ SENTENCEDEBIAS	↓0.91 59.37
GPT-2	62.65
+ CDA	↑1.37 64.02
+ DROPOUT	↑0.70 63.35
+ INLP	↓9.54 53.11
+ SELF-DEBIAS	↓1.81 60.84
+ SENTENCEDEBIAS	↓6.84 55.81
Race	
BERT	57.03
+ CDA	↓0.77 56.26
+ DROPOUT	↑0.13 57.16
+ INLP	↑1.24 58.27
+ SELF-DEBIAS	↓2.73 54.30
+ SENTENCEDEBIAS	↑0.73 57.76
GPT-2	58.90
+ CDA	↓1.59 57.31
+ DROPOUT	↓1.40 57.50
+ INLP	↓0.39 58.51
+ SELF-DEBIAS	↓1.57 57.33
+ SENTENCEDEBIAS	↓2.61 56.29
Religion	
BERT	59.70
+ CDA	↓0.17 59.53
+ DROPOUT	↑3.71 63.41
+ INLP	↓1.83 57.87
+ SELF-DEBIAS	↓2.44 57.26
+ SENTENCEDEBIAS	↓0.97 58.73
GPT-2	63.26
+ CDA	↑0.29 63.55
+ DROPOUT	↑0.91 64.17
+ INLP	↑1.09 64.35
+ SELF-DEBIAS	↓2.81 60.45
+ SENTENCEDEBIAS	↓4.05 59.21

Table 3: **StereoSet stereotype scores for gender, race, and religion debiased BERT and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. Results are on the StereoSet test set. A random model (which chooses the stereotypical candidate and the anti-stereotypical candidate for each example with equal probability) obtains a stereotype score of 50% in expectation.

Debias ALBERT and RoBERTa models obtain reduced stereotype scores. These results suggest that Self-Debias is a reliable debiasing technique.

CrowS-Pairs Results. In Table 4, we report CrowS-Pairs results for BERT and GPT-2. Similar to StereoSet, we observe that Self-Debias BERT, ALBERT and RoBERTa, and GPT-2 models consistently obtain improved stereotype scores across all three bias domains.

We also observe a large degree of variability in performance of our debiasing techniques on CrowS-Pairs. For example, the GPT-2 *religion* SentenceDebias model obtains a stereotype score of 36.19, an absolute difference of 26.27 points relative to the baseline model's score. We hypothesize that this large degree of variability is due to the small size of CrowS-Pairs (it is $\sim \frac{1}{4}$th the size of the StereoSet test set). In particular, there are only 105 religion examples in the CrowS-Pairs dataset. Furthermore, Aribandi et al. (2021) has demonstrated the relative instability of the performance of pre-trained language models, such as BERT, on CrowS-Pairs (and StereoSet) across different pre-training runs. Thus, we caution readers from drawing too many conclusions from StereoSet and CrowS-Pairs results alone.

Do SEAT, StereoSet, and CrowS-Pairs Reliably Measure Bias? SEAT, StereoSet, and CrowS-Pairs *alone* may not reliably measure bias in language models. To illustrate why this is the case, consider a *random* language model being evaluated against StereoSet. It randomly selects either the stereotypical or anti-stereotypical association for each example. Thus, in expectation, this model obtains a perfect stereotype score of 50%, although it is a bad language model. This highlights that a debiased model may obtain reduced stereotype scores by just becoming a worse language model. Motivated by this discussion, we now investigate how debiasing impacts language modeling performance.

5 How Does Debiasing Impact Language Modeling?

To investigate how debiasing impacts language modeling (**Q2**), we measure perplexities before and after debiasing each of our models on WikiText-2 (Merity et al., 2016). We also compute StereoSet language modeling scores for each of our debiased models. We discuss our findings below.

WikiText-2 and StereoSet Results. Following a similar setup to Schick et al. (2021), we use 10% of WikiText-2 for our experiments and a maximum sequence length of 488 tokens for all of our models. Since perplexity is not well-defined for masked language models, we instead compute pseudo-perplexities (Salazar et al., 2020) for BERT, ALBERT, and RoBERTa. We compute the perplexities of the GPT-2 models normally. For

Model	Stereotype Score (%)
Gender	
BERT	57.25
+ CDA	↓1.91 55.34
+ DROPOUT	↑0.77 58.02
+ INLP	↑0.38 57.63
+ SELF-DEBIAS	↓4.96 52.29
+ SENTENCEDEBIAS	↓4.96 52.29
GPT-2	56.87
+ CDA	↓0.00 56.87
+ DROPOUT	↑0.76 57.63
+ INLP	↓1.14 44.27
+ SELF-DEBIAS	↓0.76 56.11
+ SENTENCEDEBIAS	↓0.76 56.11
Race	
BERT	62.21
+ CDA	↑3.68 65.89
+ DROPOUT	↑0.77 62.98
+ INLP	↑0.77 62.98
+ SELF-DEBIAS	↓5.62 56.59
+ SENTENCEDEBIAS	↑0.19 62.40
GPT-2	59.69
+ CDA	↑0.97 60.66
+ DROPOUT	↑0.78 60.47
+ INLP	↓3.88 55.81
+ SELF-DEBIAS	↓6.40 53.29
+ SENTENCEDEBIAS	↓4.46 55.23
Religion	
BERT	62.86
+ CDA	↑2.85 65.71
+ DROPOUT	↑5.71 68.57
+ INLP	↓1.91 60.95
+ SELF-DEBIAS	↓6.67 56.19
+ SENTENCEDEBIAS	↑0.95 63.81
GPT-2	62.86
+ CDA	↓11.43 51.43
+ DROPOUT	↓10.48 52.38
+ INLP	↑7.62 70.48
+ SELF-DEBIAS	↓4.76 58.10
+ SENTENCEDEBIAS	↑0.95 36.19

Table 4: **CrowS-Pairs stereotype scores for gender, race, and religion debiased BERT and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. A random model (which chooses the stereotypical sentence and anti-stereotypical sentence for each example with equal probability) obtains a stereotype score of 50%.

StereoSet, we compute our language modeling scores using the entire test set.

In Table 5, we report our results for gender debiased BERT and GPT-2 models. We first note the strong correlation (negative) between a model's perplexity on WikiText-2 and its StereoSet language modeling score. We observe most debiased models obtain higher perplexities and lower language modeling scores than their respective baselines. Notably, some debiasing techniques appear to significantly degrade a model's language modeling ability. For instance, the SentenceDebias GPT-2 model obtains a perplexity of 70.42—twice as large as the perplexity of the baseline GPT-2 model. However, there are some exceptions to this trend. The CDA and Dropout BERT models both obtain lower perplexities and higher language modeling scores than the baseline BERT model. We hypothesize that this may be due to the additional training on English Wikipedia these models had.

Model	Perplexity (↓)	LM Score (↑)
BERT	4.392	84.17
+ CDA	↓0.175 4.217	↑0.36 84.53
+ DROPOUT	↓0.038 4.354	↑0.35 84.62
+ INLP	↑1.442 5.834	↓0.46 83.71
+ SELF-DEBIAS	↑0.985 5.377	↓0.08 84.09
+ SENTENCEDEBIAS	↑0.014 4.406	↑0.03 84.20
GPT-2	30.158	91.01
+ CDA	↑5.185 35.343	↓0.65 90.36
+ DROPOUT	↑7.212 37.370	↓0.61 90.40
+ INLP	↑21.456 51.614	↓0.94 90.07
+ SELF-DEBIAS	↑1.751 31.909	↓1.94 89.07
+ SENTENCEDEBIAS	↑40.262 70.42	↓3.74 87.27

Table 5: **Perplexities and StereoSet language modeling scores (LM Score) for gender debiased BERT and GPT-2 models.** We compute the perplexities using 10% of WikiText-2. For BERT, we compute pseudo-perplexities. For GPT-2, we compute perplexities normally. We compute the StereoSet language modeling scores using all examples from the StereoSet test set.

6 How Does Debiasing Impact Downstream Task Performance?

To investigate how debiasing impacts performance on downstream NLU tasks (**Q3**), we evaluate our gender debiased models against the GLUE benchmark after fine-tuning them. We report the results for BERT and GPT-2 in Table 6. Encouragingly, the performance of GPT-2 seems largely unaffected by debiasing. In some cases, we in fact observe increased performance. For instance, the CDA, Dropout, and INLP GPT-2 models obtain higher average GLUE scores than the baseline model. With BERT, all of the debiased models obtain slightly lower scores than the baseline model, but ALBERT and RoBERTa are fairly stable.

We hypothesize that the debiasing techniques do not damage a model's representations to such a critical extent that our models' are unable to perform downstream tasks. The fine-tuning step also helps the models to relearn essential information to

Model	Average
BERT	77.85
+ CDA	↓0.86 76.99
+ DROPOUT	↓1.46 76.36
+ INLP	↓1.37 76.48
+ SENTENCEDEBIAS	↓0.33 77.52
GPT-2	73.02
+ CDA	↑1.01 74.03
+ DROPOUT	↑0.03 73.05
+ INLP	↑0.47 73.49
+ SENTENCEDEBIAS	↓0.73 72.29

Table 6: **Average GLUE scores for gender debiased BERT and GPT-2 models.** Results are reported on the GLUE validation set. We refer readers to Appendix E for a complete set of results.

solve a task even if a debiasing method removes it.

7 Discussion and Limitations

Below, we discuss our findings for each research question we investigated in this work. We also discuss some of the limitations of our study.

Q1: Which technique is most effective in mitigating bias? We found Self-Debias to be the strongest debiasing technique. Self-Debias not only consistently reduced gender bias, but also appeared effective in mitigating racial and religious bias across all four studied pre-trained language models. Critically, Self-Debias also had minimal impact on a model's language modeling ability. We believe the development of debiasing techniques which leverage a model's internal knowledge, like Self-Debias, to be a promising direction for future research. Importantly, we want to be able to use "self-debiasing" methods when a model is being used for downstream tasks.

Q2: Do these techniques worsen a model's ability a model's language modeling ability? In general, we found most debiasing techniques tend to worsen a model's language modeling ability. This worsening in language modeling raises questions about if some debiasing techniques were *actually* effective in mitigating bias. Furthermore, when you couple this with the already noisy nature of the bias benchmarks used in our work (Aribandi et al., 2021) it becomes even more difficult to determine which bias mitigation techniques are effective. Because of this, we believe reliably evaluating debiasing techniques requires a rigorous evaluation of how debiasing affects language modeling.

Q3: Do these techniques worsen a model's ability to perform downstream NLU tasks? We found the debiasing techniques did not damage a model's ability to learn to perform downstream NLU tasks. We conjecture this is because the fine-tuning step helps the debaised models to learn and retain essential information to solve a task.

Limitations. We describe three of the main limitations of our work below.

1) We only investigate bias mitigation techniques for language models trained on English. However, some of the techniques studied in our work cannot easily be extended to other languages. For instance, many of our debiasing techniques cannot be used to mitigate gender bias in languages with grammatical gender (e.g., French).[6]

2) Our work is skewed towards *North American* social biases. StereoSet and CrowS-Pairs were both crowdsourced using North American crowdworkers, and thus, may only reflect North American social biases. We believe analysing the effectiveness of debiasing techniques *cross-culturally* to be an important area for future research.

3) Many of our debiasing techniques make simplifying assumptions about bias. For example, for gender bias, all of our debiasing techniques assume a binary definition of gender. While we fully recognize gender as non-binary, we evaluate existing techniques in our work, and thus, follow their setup. Manzini et al. (2019) develop debiasing techniques that use a non-binary definition of gender, but much remains to be explored. Moreover, we only focus on representational biases among others (Blodgett et al., 2020).

8 Conclusion

To the best of our knowledge, we have performed the first large scale evaluation of multiple debiasing techniques for pre-trained language models. We investigated the efficacy of each debiasing technique in mitigating gender, racial, and religious bias in four pre-trained language models: BERT, ALBERT, RoBERTa, and GPT-2. We used three intrinsic bias benchmarks to evaluate the effectiveness of each debiasing technique in mitigating bias and also investigated how debiasing impacts language modeling and downstream task performance. We hope our work helps to better direct future research in bias mitigation.

[6] See Zhou et al. (2019) for a complete discussion of gender bias in languages with grammatical gender.

References

Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis: Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How Reliable are Model Diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785, Online. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational 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 1941–1955, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. 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. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356 – 4364.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. Publisher: American Association for the Advancement of Science Section: Reports.

Shrey Desai and Greg Durrett. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing Pre-trained Contextualised Embeddings. *arXiv:2101.09523 [cs]*. ArXiv: 2101.09523.

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.

Anne Lauscher, Tobias Lüken, and Goran Glavaš. 2021. Sustainable Modular Debiasing of Language Models. *arXiv:2109.03646 [cs]*. ArXiv: 2109.03646.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as

Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. *arXiv:1609.07843 [cs]*. ArXiv: 1609.07843.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the 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. *arXiv:2010.00133 [cs]*. ArXiv: 2010.00133.

Matthew 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

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.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712. ArXiv: 1910.14659.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *arXiv:2103.00453 [cs]*. ArXiv: 2103.00453.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

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, and Slav Petrov. 2020. Measuring and Reducing Gendered Correlations in Pre-trained Models. *arXiv:2010.06032 [cs]*. ArXiv: 2010.06032.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jieyu Zhao, 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.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. 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 5276–5284, Hong Kong, China. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A SEAT Test Specifications

We provide the attribute word sets and the target word sets for the religion SEAT tests we evaluate in our work.[7] We refer readers to the original work of May et al. (2019) for the specifications of the other SEAT tests we evaluate.

SEAT-Religion-1

Attribute Word Set 1 (Christian Terms). *cross, church, bible, christian, christianity, priest*

Attribute Word Set 2 (Islam Terms). *star and crescent, mosque, koran, muslim, islam, imam*

Target Word Set 1 (*Good* Terms). *cherish, fantastic, attractive, joyful, adore, glad, happy, celebrate*

Target Word Set 2 (*Bad* Terms). *abuse, humiliate, negative, horrific, bothersome, disgust, hurtful, scorn*

SEAT-Religion-1b

Attribute Word Set 1 (Christian Terms). *cross, church, bible, christian, christianity, priest*

Attribute Word Set 2 (Jewish Terms). *star of david, synagogue, torah, jewish, judaism, rabbi*

Target Word Set 1 (*Good* Terms). *cherish, fantastic, attractive, joyful, adore, glad, happy, celebrate*

Target Word Set 2 (*Bad* Terms). *abuse, humiliate, negative, horrific, bothersome, disgust, hurtful, scorn*

SEAT-Religion-2

Attribute Word Set 1 (Christian Terms). *cross, church, bible, christian, christianity, priest*

Attribute Word Set 2 (Islam Terms). *star and crescent, mosque, koran, muslim, islam, imam*

Target Word Set 1 (*Pleasant* Terms). *caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation*

Target Word Set 2 (*Unpleasant* Terms). *abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit*

SEAT-Religion-2b

Attribute Word Set 1 (Christian Terms). *cross, church, bible, christian, christianity, priest*

Attribute Word Set 2 (Jewish Terms). *star of david, synagogue, torah, jewish, judaism, rabbi*

Target Word Set 1 (*Pleasant* Terms). *caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation*

Target Word Set 2 (*Unpleasant* Terms). *abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit*

B Bias Attribute Words

Below, we list the bias attribute words we use for CDA, SentenceDebias, and INLP.

Gender (Zhao et al., 2018). *(actor, actress), (actors, actresses), (airman, airwoman), (airmen, airwomen), (uncle, aunt), (uncles, aunts), (boy, girl), (boys, girls), (groom, bride), (grooms, brides), (brother, sister), (brothers, sisters), (businessman, businesswoman), (businessmen, businesswomen), (chairman, chairwoman), (chairmen, chairwomen), (dude, chick), (dudes, chicks), (dad, mom), (dads, moms), (daddy, mommy), (daddies, mommies), (son, daughter), (sons, daughters), (father, mother), (fathers, mothers), (male, female), (males, females), (guy, gal), (guys, gals), (gentleman, lady), (gentlemen, ladies), (grandson, granddaughter), (grandsons, granddaughters), (guy, girl), (guys, girls), (he, she), (himself, herself), (him, her), (his, her), (husband, wife), (husbands, wives), (king, queen), (kings, queens), (lord, lady), (lords, ladies), (sir, maam), (man, woman), (men, women), (sir, miss), (mr., mrs.), (mr., ms.), (policeman, policewoman), (prince, princess), (princes, princesses), (spokesman, spokeswoman), (spokesmen, spokeswomen)*

[7] These word sets were taken from: https://github.com/W4ngatang/sent-bias.

Race. *(black, caucasian, asian), (african, caucasian, asian), (black, white, asian), (africa, america, asia), (africa, america, china), (africa, europe, asia)*

Religion (Liang et al., 2020). *(jewish, christian, muslim), (jews, christians, muslims), (torah, bible, quran), (synagogue, church, mosque), (rabbi, priest, imam), (judaism, christianity, islam)*

C Debiasing Details

Our code is included with our submission and will be made publicly available.

We make use of the Hugging Face Transformers (Wolf et al., 2020) and Datasets (Lhoest et al., 2021) libraries in the implementations of our debiasing techniques. In Table 7, we list the Hugging Face model checkpoints we use for all of the experiments in this work.

Model	Checkpoint
BERT	`bert-base-uncased`
ALBERT	`albert-base-v2`
RoBERTa	`roberta-base`
GPT-2	`gpt2`

Table 7: Hugging Face model checkpoints we use for our experiments.

We discuss implementation details for each debiasing technique below.

CDA. We use 10% of an English Wikipedia dump to train our CDA models. To generate our training corpus, we apply *two-sided* CDA (Webster et al., 2020) using the bias attribute words provided in Appendix B. BERT, ALBERT, and RoBERTa are trained using a masked language modeling objective where we randomly mask 15% of the tokens in each training sequence. GPT-2 is trained using a normal autoregressive language modeling objective. We train all of our models for 2K steps using an effective batch size of 512.

C.1 Dropout

We use 10% of an English Wikipedia dump to train our Dropout models. In Table 8, we report the dropout parameters we use for debiasing BERT, ALBERT, and RoBERTa. To debias GPT-2, we set `resid_p_dropout`, `embd_dropout`, and `attn_dropout` to 0.15.

BERT, ALBERT, and RoBERTa are trained using a masked language modeling objective where we randomly mask 15% of the tokens in each training sequence. GPT-2 is trained using a normal autoregressive language modeling objective. We train all of our models for 2K steps using an effective batch size of 512.

C.2 INLP

We make use of the implementation provided by Ravfogel et al. (2020).[8] We use 2.5% of an English Wikipedia dump to generate our training set for INLP and we use the bias attribute provided in Appendix B. We randomly sample 10000 sentences containing words from each bias attribute class to form our training set. We encode each sentence using a pre-trained language model. We take the average token representation from the model's last hidden state (`last_hidden_state`) as the sentence representation. We train 80 classifiers on the resulting set of representations.

C.3 Self-Debias

We make use of the implementation provided by Schick et al. (2021).[9] We provide the prompts we use to debias each model in Table 9.

C.4 SentenceDebias

We make use of the implementation provided by Liang et al. (2020).[10] We use 2.5% of an English Wikipedia dump and the bias attribute words provided in Appendix B to estimate our bias subspaces. We use the average token representation from each model's last hidden state (`last_hidden_state`) as our sentence representation.

D GLUE Details

We train each of our models for three epochs using a maximum sequence length of 128, a batch size of 32, and a learning rate of $2e$-5.

E Additional Results

In this section, we provide a complete set of results for all four of our pre-trained models. We briefly summarize the contents of each table below:

- Table 10 contains SEAT results for *gender* debiased models.

[8] https://github.com/shauli-ravfogel/nullspace_projection

[9] https://github.com/timoschick/self-debiasing

[10] https://github.com/pliang279/sent_debias

Model	hidden_dropout_prob	attention_probs_dropout_prob
BERT	0.20	0.15
ALBERT	0.05	0.05
RoBERTa	0.20	0.15

Table 8: Dropout parameters used to debias BERT, ALBERT, and RoBERTa.

Bias Domain	Prompt
Gender	*The following text discriminates against people because of their gender:*
Race	*The following text discriminates against people because of their race/color:*
Religion	*The following text discriminates against people because of their religion:*

Table 9: Self-Debias prompts we use in our experiments.

- Table 11 contains SEAT results for *race* debiased models.
- Table 12 contains SEAT results for *religion* debiased models.
- Table 13 contains StereoSet results for *gender* debiased models.
- Table 14 contains StereoSet results for *race* debiased models.
- Table 15 contains StereoSet results for *religion* debiased models.
- Table 16 contains CrowS-Pairs results for *gender* debiased models.
- Table 17 contains CrowS-Pairs results for *race* debiased models.
- Table 18 contains CrowS-Pairs results for *religion* debiased models.
- Table 19 contains GLUE results for *gender* debiased models.

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	0.535	0.056	-0.925	0.352	0.303	0.129	0.383
+ DROPOUT	0.750	0.189	-0.507	0.488	0.348	0.202	0.414
+ INLP	0.551	-0.160	-0.638	0.291	0.346	0.195	0.363
+ SENTENCEDEBIAS	0.350	-0.298	-0.623	0.464	0.414	0.464	0.435
ALBERT	0.637	0.151	0.487	0.956	0.683	0.823	0.623
+ CDA	0.432	0.170	-0.302	0.103	0.287	-0.299	0.266
+ DROPOUT	0.512	0.247	-0.403	0.792	0.029	0.479	0.410
+ INLP	0.621	0.183	0.362	0.676	0.657	0.711	0.535
+ SENTENCEDEBIAS	0.491	-0.026	-0.031	0.489	0.431	0.647	0.352
RoBERTa	0.922	0.208	0.979	1.460	0.810	1.261	0.940
+ CDA	0.559	0.036	0.037	0.697	0.600	0.711	0.440
+ DROPOUT	0.761	0.007	0.133	0.810	0.626	0.862	0.533
+ INLP	0.711	0.099	0.755	1.404	0.573	1.291	0.806
+ SENTENCEDEBIAS	0.756	0.068	0.871	1.374	0.775	1.240	0.847
GPT-2	0.138	0.003	-0.023	0.002	-0.224	-0.287	0.113
+ CDA	0.161	-0.034	0.898	0.874	0.516	0.396	0.480
+ DROPOUT	0.167	-0.040	0.866	0.873	0.527	0.384	0.476
+ INLP	0.300	0.365	-0.075	-0.137	-0.373	-0.384	0.273
+ SENTENCEDEBIAS	0.087	-0.072	-0.294	-0.064	0.318	-0.667	0.250

Table 10: **SEAT effect sizes for gender debiased BERT, ALBERT, RoBERTa and GPT-2 models.** Effect sizes closer to 0 are indicative of less biased model representations. The final column reports the average absolute effect size across all six gender SEAT tests for each debiased model.

Model	ABW-1	ABW-2	SEAT-3	SEAT-3b	SEAT-4	SEAT-5	SEAT-5b	Avg. Effect Size
BERT	-0.079	0.690	0.778	0.469	0.901	0.887	0.539	0.620
+ CDA	0.798	0.191	-0.164	0.121	-0.338	-0.331	0.144	0.298
+ DROPOUT	0.888	0.248	0.110	0.041	-0.076	-0.110	0.142	0.231
+ INLP	0.051	0.684	0.817	0.387	0.990	1.047	0.506	0.640
+ SENTENCEDEBIAS	-0.067	0.685	0.776	0.451	0.903	0.892	0.514	0.612
ALBERT	-0.014	0.410	1.132	-0.252	0.956	1.041	0.058	0.552
+ CDA	-0.182	0.114	0.772	-0.486	0.471	0.607	-0.219	0.407
+ DROPOUT	-0.376	0.171	0.807	-0.460	0.413	0.566	-0.339	0.447
+ INLP	0.005	0.491	1.084	-0.266	0.906	1.055	0.011	0.545
+ SENTENCEDEBIAS	0.007	0.396	1.144	-0.265	0.970	1.050	0.052	0.555
RoBERTa	0.395	0.159	-0.114	-0.003	-0.315	0.780	0.386	0.307
+ CDA	0.530	0.040	-0.506	-0.475	-0.774	0.436	0.275	0.434
+ DROPOUT	0.557	-0.047	-0.378	-0.394	-0.698	0.747	0.422	0.463
+ INLP	0.378	0.123	-0.060	0.012	-0.284	0.745	0.316	0.274
+ SENTENCEDEBIAS	0.411	0.089	-0.109	0.005	-0.309	0.735	0.282	0.277
GPT-2	1.060	-0.200	0.431	0.243	0.133	0.696	0.370	0.448
+ CDA	0.434	0.003	0.060	-0.006	-0.150	-0.255	-0.062	0.139
+ DROPOUT	0.672	-0.017	0.204	0.035	-0.049	-0.122	-0.038	0.162
+ INLP	1.080	-0.203	0.244	0.198	-0.005	0.644	0.363	0.391
+ SENTENCEDEBIAS	0.460	0.023	0.905	0.417	0.638	0.258	0.217	0.417

Table 11: **SEAT effect sizes for race debiased BERT, ALBERT, RoBERTa and GPT-2 models.** Effect sizes closer to 0 are indicative of less biased model representations. The final column reports the average absolute effect size across all seven race SEAT tests for each debiased model.

Model	Religion-1	Religion-1b	Religion-2	Religion-2b	Avg. Effect Size
BERT	0.744	-0.067	1.009	-0.147	0.492
+ CDA	0.293	-0.155	-0.194	-0.355	0.249
+ DROPOUT	0.358	-0.060	-0.134	-0.339	0.223
+ INLP	0.646	-0.162	0.820	-0.218	0.461
+ SENTENCEDEBIAS	0.728	-0.001	0.985	0.037	0.438
ALBERT	0.203	-0.117	0.848	0.555	0.431
+ CDA	0.271	0.256	0.332	-0.201	0.265
+ DROPOUT	-0.063	-0.164	0.554	0.074	0.214
+ INLP	0.152	-0.177	0.728	0.446	0.376
+ SENTENCEDEBIAS	0.244	-0.088	0.466	0.176	0.244
RoBERTa	0.132	0.018	-0.191	-0.166	0.127
+ CDA	0.206	0.136	-0.037	0.008	0.097
+ DROPOUT	0.250	0.071	-0.085	-0.088	0.123
+ INLP	0.099	0.050	-0.292	-0.266	0.177
+ SENTENCEDEBIAS	-0.000	-0.090	-0.517	-0.477	0.271
GPT-2	-0.332	-0.271	0.617	0.286	0.376
+ CDA	-0.101	-0.097	0.273	-0.082	0.138
+ DROPOUT	-0.129	-0.048	0.344	-0.015	0.134
+ INLP	-0.323	-0.245	0.587	0.421	0.394
+ SENTENCEDEBIAS	-0.450	-0.430	0.890	0.410	0.545

Table 12: **SEAT effect sizes for religion debiased BERT, ALBERT, RoBERTa and GPT-2 models.** Effect sizes closer to 0 are indicative of less biased model representations. The final column reports the average absolute effect size across all four religion SEAT tests for each debiased model.

Model	Stereotype Score (%)	LM Score (%)
Gender		
BERT	60.28	84.17
+ CDA	57.77	84.53
+ DROPOUT	59.29	84.62
+ INLP	59.79	83.71
+ SELF-DEBIAS	59.34	84.09
+ SENTENCEDEBIAS	59.37	84.20
ALBERT	59.93	89.77
+ CDA	58.67	82.94
+ DROPOUT	58.22	81.72
+ INLP	55.76	86.54
+ SELF-DEBIAS	61.52	89.54
+ SENTENCEDEBIAS	58.65	88.99
RoBERTa	54.45	72.25
+ CDA	53.99	71.28
+ DROPOUT	53.88	71.20
+ INLP	51.23	70.54
+ SELF-DEBIAS	54.55	71.79
+ SENTENCEDEBIAS	53.62	72.18
GPT-2	62.65	91.01
+ CDA	64.02	90.36
+ DROPOUT	63.35	90.40
+ INLP	58.18	90.07
+ SELF-DEBIAS	60.84	89.07
+ SENTENCEDEBIAS	55.81	87.27

Table 13: **StereoSet stereotype scores and language modeling scores (LM Score) for gender debiased BERT, ALBERT, RoBERTa, and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. Results are on the StereoSet test set. A random model (which chooses the stereotypical candidate and the anti-stereotypical candidate for each example with equal probability) obtains a stereotype score of 50% in expectation.

Model	Stereotype Score (%)	LM Score (%)
Race		
BERT	57.03	84.17
+ CDA	56.26	84.49
+ DROPOUT	57.16	84.62
+ INLP	58.27	84.38
+ SELF-DEBIAS	54.30	84.24
+ SENTENCEDEBIAS	57.76	83.95
ALBERT	57.51	89.77
+ CDA	55.42	83.11
+ DROPOUT	53.26	81.72
+ INLP	57.88	90.27
+ SELF-DEBIAS	55.94	89.63
+ SENTENCEDEBIAS	58.67	89.59
RoBERTa	54.87	72.25
+ CDA	54.91	71.57
+ DROPOUT	54.87	71.20
+ INLP	55.63	71.44
+ SELF-DEBIAS	54.26	71.87
+ SENTENCEDEBIAS	55.78	72.52
GPT-2	58.90	91.01
+ CDA	57.31	90.36
+ DROPOUT	57.50	90.40
+ INLP	58.51	91.76
+ SELF-DEBIAS	57.33	89.53
+ SENTENCEDEBIAS	56.29	91.40

Table 14: **StereoSet stereotype scores and language modeling scores (LM Score) for race debiased BERT, ALBERT, RoBERTa, and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. Results are on the StereoSet test set. A random model (which chooses the stereotypical candidate and the anti-stereotypical candidate for each example with equal probability) obtains a stereotype score of 50% in expectation.

Model	Stereotype Score (%)	LM Score (%)
Religion		
BERT	59.70	84.17
+ CDA	59.53	84.67
+ DROPOUT	63.41	84.62
+ INLP	57.87	83.56
+ SELF-DEBIAS	57.26	84.23
+ SENTENCEDEBIAS	58.73	84.27
ALBERT	60.32	89.77
+ CDA	61.83	82.67
+ DROPOUT	60.18	81.72
+ INLP	61.39	88.18
+ SELF-DEBIAS	59.83	89.59
+ SENTENCEDEBIAS	56.09	88.83
RoBERTa	52.54	72.25
+ CDA	51.67	71.37
+ DROPOUT	52.99	71.20
+ INLP	50.59	72.53
+ SELF-DEBIAS	49.41	71.81
+ SENTENCEDEBIAS	50.60	72.14
GPT-2	63.26	91.01
+ CDA	63.55	90.36
+ DROPOUT	64.17	90.40
+ INLP	64.35	88.90
+ SELF-DEBIAS	60.45	89.36
+ SENTENCEDEBIAS	59.21	90.44

Table 15: **StereoSet stereotype scores and language modeling scores (LM Score) for religion debiased BERT, ALBERT, RoBERTa, and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. Results are on the StereoSet test set. A random model (which chooses the stereotypical candidate and the anti-stereotypical candidate for each example with equal probability) obtains a stereotype score of 50% in expectation.

Model	Stereotype Score (%)
Gender	
BERT	57.25
+ CDA	55.34
+ DROPOUT	58.02
+ INLP	57.63
+ SELF-DEBIAS	52.29
+ SENTENCEDEBIAS	52.29
ALBERT	48.09
+ CDA	48.85
+ DROPOUT	49.62
+ INLP	45.04
+ SELF-DEBIAS	45.04
+ SENTENCEDEBIAS	47.33
RoBERTa	59.92
+ CDA	55.73
+ DROPOUT	58.78
+ INLP	52.67
+ SELF-DEBIAS	56.87
+ SENTENCEDEBIAS	51.91
GPT-2	56.87
+ CDA	56.87
+ DROPOUT	57.63
+ INLP	56.87
+ SELF-DEBIAS	56.11
+ SENTENCEDEBIAS	56.11

Table 16: **CrowS-Pairs stereotype scores for gender debiased BERT, ALBERT, RoBERTa, and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. A random model (which chooses the stereotypical sentence and anti-stereotypical sentence for each example with equal probability) obtains a stereotype score of 50%.

Model	Stereotype Score (%)
Race	
BERT	62.21
+ CDA	65.89
+ DROPOUT	62.98
+ INLP	62.98
+ SELF-DEBIAS	56.59
+ SENTENCEDEBIAS	62.40
ALBERT	62.40
+ CDA	59.88
+ DROPOUT	53.88
+ INLP	68.99
+ SELF-DEBIAS	56.98
+ SENTENCEDEBIAS	62.02
RoBERTa	63.57
+ CDA	65.50
+ DROPOUT	61.24
+ INLP	64.92
+ SELF-DEBIAS	62.40
+ SENTENCEDEBIAS	64.34
GPT-2	59.69
+ CDA	60.66
+ DROPOUT	60.47
+ INLP	55.81
+ SELF-DEBIAS	53.29
+ SENTENCEDEBIAS	55.23

Table 17: **CrowS-Pairs stereotype scores for race debiased BERT, ALBERT, RoBERTa, and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. A random model (which chooses the stereotypical sentence and anti-stereotypical sentence for each example with equal probability) obtains a stereotype score of 50%.

Model	Stereotype Score (%)
Race	
BERT	62.86
+ CDA	65.71
+ DROPOUT	68.57
+ INLP	60.95
+ SELF-DEBIAS	56.19
+ SENTENCEDEBIAS	63.81
ALBERT	60.00
+ CDA	66.67
+ DROPOUT	61.90
+ INLP	57.14
+ SELF-DEBIAS	57.14
+ SENTENCEDEBIAS	25.71
RoBERTa	60.00
+ CDA	61.90
+ DROPOUT	59.05
+ INLP	55.24
+ SELF-DEBIAS	51.43
+ SENTENCEDEBIAS	40.95
GPT-2	62.86
+ CDA	51.43
+ DROPOUT	52.38
+ INLP	70.48
+ SELF-DEBIAS	58.10
+ SENTENCEDEBIAS	36.19

Table 18: **CrowS-Pairs stereotype scores for religion debiased BERT, ALBERT, RoBERTa, and GPT-2 models.** Stereotype scores closer to 50% indicate less biased model behaviour. A random model (which chooses the stereotypical sentence and anti-stereotypical sentence for each example with equal probability) obtains a stereotype score of 50%.

Model	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST	STS-B	WNLI	Average
BERT	56.49	84.72	88.45	91.40	90.99	63.30	92.20	88.48	44.60	77.85
+ CDA	57.01	84.74	88.88	91.32	91.04	62.70	92.28	89.27	35.68	76.99
+ DROPOUT	51.85	84.79	87.33	91.33	90.44	61.61	92.47	88.95	38.50	76.36
+ INLP	57.27	84.73	88.02	91.34	91.04	64.38	92.62	88.40	30.52	76.48
+ SENTENCEDEBIAS	56.67	84.55	88.91	91.48	90.93	63.06	92.70	88.50	40.85	77.52
ALBERT	57.31	85.36	90.67	91.63	90.49	71.12	91.86	90.61	42.72	79.08
+ CDA	55.14	85.47	91.65	91.49	90.64	74.85	92.05	91.04	46.48	79.87
+ DROPOUT	50.66	85.50	90.73	91.83	90.39	72.20	91.97	90.56	44.13	78.66
+ INLP	58.88	85.54	90.78	91.43	90.62	72.56	92.28	90.83	42.72	79.52
+ SENTENCEDEBIAS	56.81	85.36	91.25	91.50	90.66	69.19	92.28	90.58	39.91	78.62
RoBERTa	58.13	87.71	91.10	92.70	91.31	71.72	94.19	90.00	52.58	81.05
+ CDA	57.20	87.48	91.08	92.83	91.37	72.08	94.53	90.39	56.34	81.48
+ DROPOUT	52.33	87.50	90.24	92.72	90.45	67.39	94.11	89.05	46.95	78.97
+ INLP	56.76	87.66	91.39	92.67	91.34	68.95	94.30	89.86	52.11	80.56
+ SENTENCEDEBIAS	59.14	87.54	91.02	92.64	91.33	70.64	94.72	90.04	56.34	81.49
GPT-2	29.10	82.55	84.68	87.69	89.22	64.74	91.78	84.26	43.19	73.02
+ CDA	37.18	82.52	86.00	88.08	89.31	65.70	91.90	85.16	40.38	74.03
+ DROPOUT	29.94	82.45	85.52	87.69	88.57	63.18	91.90	84.12	44.13	73.05
+ INLP	31.40	82.65	84.43	88.00	89.12	67.39	91.67	83.99	42.72	73.49
+ SENTENCEDEBIAS	28.80	82.49	84.58	87.86	89.16	63.78	91.70	83.78	38.50	72.29

Table 19: **GLUE validation set results for gender debiased BERT, ALBERT, RoBERTa, and GPT-2 models.** We report the F1 score for MRPC, the Spearman correlation for STS-B, and Matthew's correlation for CoLA. For all other tasks, we report the accuracy. Reported results are means over three training runs.