This item is the archived preprint of:

How far can it go? On intrinsic gender bias mitigation for text classification

How Far Can It Go? On Intrinsic Gender Bias Mitigation for Text Classification

Ewoenam Kwaku Tokpo[1]**, Pieter Delobelle**[2,3,5]**, Bettina Berendt**[2,3,4,5] and **Toon Calders**[1]

[1] Department of Computer Science, University of Antwerp, Belgium
[2] Department of Computer Science, KU Leuven, Belgium
[3] Leuven.AI Institute for Artificial Intelligence, Belgium
[4] Faculty of Electrical Engineering and Computer Science, TU Berlin, Germany
[5] Weizenbaum Institute, Germany

Abstract

To mitigate gender bias in contextualized language models, different intrinsic mitigation strategies have been proposed, alongside many bias metrics. Considering that the end use of these language models is for downstream tasks like text classification, it is important to understand how these intrinsic bias mitigation strategies actually translate to fairness in downstream tasks and the extent of this. In this work, we design a probe to investigate the effects that some of the major intrinsic gender bias mitigation strategies have on downstream text classification tasks. We discover that instead of resolving gender bias, intrinsic mitigation techniques and metrics are able to hide it in such a way that significant gender information is retained in the embeddings. Furthermore, we show that each mitigation technique is able to hide the bias from some of the intrinsic bias measures but not all, and each intrinsic bias measure can be fooled by some mitigation techniques, but not all. We confirm experimentally, that none of the intrinsic mitigation techniques used without any other fairness intervention is able to consistently impact extrinsic bias. We recommend that intrinsic bias mitigation techniques should be combined with other fairness interventions for downstream tasks.

1 Introduction

The use of pretrained language models has seen a surge in popularity as a result of state-of-the-art performances that have been achieved with these models on various tasks. Consequently, there has been a growing interest in exploring how gender bias pertains in these models (Garrido-Muñoz et al., 2021). Pretrained language models are used in two distinct phases: the *pretraining phase* and the *finetuning phase*. The pretraining phase typically involves training a model on a generic task such as masked language modeling on a diverse set of text corpora. In the finetuning phase, the pretrained model can be adapted for a specific task, such as sentiment analysis, by training on a domain-specific corpus. Owing to the unique way of using pretrained models, bias generally manifests in two forms: *intrinsic bias* and *extrinsic bias*. Intrinsic bias refers to bias that inherently exists in pretrained language models whereas extrinsic bias is used to refer to bias that exists in downstream models that are based on the pretrained model. Since the success of downstream NLP tasks has mostly been dependent on pretrained models, it is intuitive to assume that bias in intrinsic models will compromise fairness in downstream tasks. Only recently have more in-depth examinations been done to investigate this assumption (Steed et al., 2022; Orgad et al., 2022; Kurita et al., 2019). However, conclusions have been inconsistent and the confounding effects of bias mitigation techniques remain unknown.

The main focus of this work is to investigate the impact of intrinsic bias on extrinsic fairness and if techniques to mitigate intrinsic bias actually resolve bias or only mask it. We develop a probe to uncover intrinsic bias by determining the amount of gender information in a word embedding using a classifier. Bearing in mind how abstractly and improperly intrinsic bias has been defined (Blodgett et al., 2020), coupled with the discovery that results from different metrics for intrinsic bias in many cases do not correlate (Delobelle et al., 2022), we find this probe effective as an extra step in evaluating the efficacy of these mitigation strategies. We realize from this study that how intrinsic bias has been measured and the choice of bias mitigation strategies explored by some existing works have not been ideal. We further use this probe to investigate if some proposed mitigation strategies superficially conceal bias.

In this work, when we refer to bias in a language model, we mean stereotyping bias as defined by Garrido-Muñoz et al. (2021) as *"the undesired variation of the [association] of certain words in that language model according to certain prior words*

in a given domain". We focus our experiments on gender bias for two primary reasons: its intuitive nature making it easy to analyze and discuss, and the accessibility of resources and datasets regarding gender. For the same reason, we narrow our definition of bias in our experiments and analysis to binary gender bias[1]. This paper considers the primary goal of mitigating intrinsic bias to ensure fairness in downstream tasks.

We consider one case of binary classification and one multiclass classification case, all on English language corpora using the BERT-large (Devlin et al., 2019) and the ALBERT-large (Lan et al., 2019) pretrained models for each task.

In summary, we develop an extensive probe [2] to uncover intrinsic bias in pertained contextualized language models, and seek to answer three key research questions: **RQ1:** Do different intrinsic bias metrics respond differently to different bias mitigation techniques? (§ 4.1). **RQ2:** Can intrinsic bias mitigation techniques hide bias instead of resolving it? (§ 5.1). **RQ3:** Do intrinsic bias mitigation techniques in language models improve fairness in downstream tasks? (§ 5.2)

2   Measuring and mitigating bias

Since many techniques for measuring and mitigating gender bias have been introduced for both intrinsic and extrinsic bias, we only discuss techniques we use in experiments in this paper.

Bias mitigation techniques can be applied to pretrained or finetuned language models, or both. Figure 1 illustrates these interactions with both training settings and this section will discuss both intrinsic (§ 2.1) and extrinsic (§ 2.3) mitigation techniques. Additionally, this section will provide a brief overview of bias measures [3].

2.1   Intrinsic bias mitigation techniques

Intrinsic gender bias mitigation methods target either the pretraining data, the pretraining procedure, or the pretrained model's output, which we refer to as pre-processing, in-processing, and post-processing respectively (Friedler et al., 2019). We select three popular mitigation methods to represent all three types, namely *Counterfactual Data*

Augmentation (CDA), *Context-debias*, and *Sent-debias*. Notice that these methods create debiased pretrained language models, as is illustrated in Figure 1. These models still need to be finetuned on a downstream task.

CDA pretraining.   The idea behind counterfactual data augmentation (Zmigrod et al., 2019; Lu et al., 2020) is to generate a counterfactual for each example in the training corpus by replacing attribute terms with their complimentary equivalent from the other demographic classes. For example, *she* will map to *he* in the case of binary gender. To mitigate intrinsic bias, this counterfactual augmentation has to be done as a pretraining step. Since CDA involves retraining the model, it is more resource-intensive compared to Sent-debias and Context-debias. We use the pretrained CDA models based on BERT and ALBERT from Webster et al. (2020) for our implementation.

Context-debias.   Kaneko and Bollegala (2021) introduce a debiasing method that involves retraining the language model with a constraint to make the embeddings of *stereotype terms* [4] (such as *doctor, nurse*) orthogonal to embeddings of *attribute terms* [5] (such as gender pronouns like *she, he* and gender nouns like *woman, man*). Given the dynamic nature of contextualized word embeddings which causes a word to have different embeddings, they define fixed word embeddings for each attribute token by averaging the contextual embeddings of a word in all sentences it appears in. Training is done so that the embeddings of all stereotype terms are made orthogonal to all fixed attribute embeddings. They add a regularizer that constrains the debiased embedding to retain as much information by ensuring that they are as similar to the original embeddings as possible despite the orthogonality (§ B.1). Context-debias, as well as other in-processing techniques, require a predefined set of attribute and target terms before training which may not be effective for words outside these sets.

Sent-debias.   Liang et al. (2020) propose a post-processing debiasing method akin to *word embedding debiasing* (Bolukbasi et al., 2016) but for contextualized embeddings. They achieve this by first identifying the bias subspace. They extract naturally-occurring sentences from corpora that

[4]We also refer to them as *target words*.
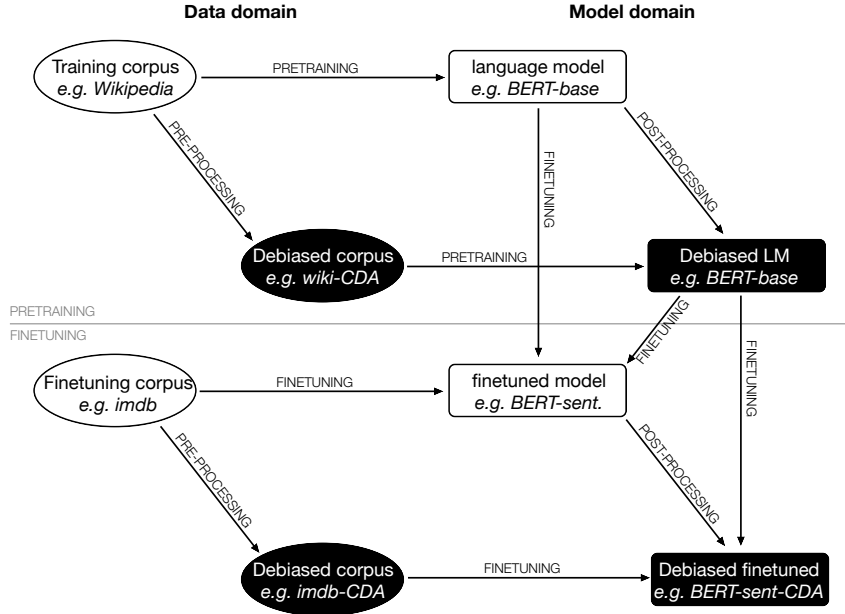[5]We also refer to them as *identity terms* or *protected words*.

Figure 1: Illustration of where bias mitigation techniques alter a model or training data for both intrinsic (pretrained) and extrinsic (finetuned) biases in models.

contain certain attribute terms, generate counterfactuals for each one, and compute the gender subspace based on these sentence templates. The rest of their algorithm follows the method of hard debiasing (neutralization and equalization) by Bolukbasi et al. (2016) (§ B.2). Essentially, this post-hoc approach transforms the embeddings by removing their projections onto the gender subspace.

2.2 Measuring intrinsic bias

To measure intrinsic bias in the pretrained models, we use *SEAT* and *LPBS*. We examine these two metrics because of their wide use in related literature.

Sentence Embedding Association Test (SEAT). May et al. (2019) developed SEAT to quantify intrinsic bias in contextualized language models based on WEAT (Caliskan et al., 2017) which was originally designed for non-contextualized embeddings. Although the authors express concern over the efficacy of this metric, it has been widely adopted in various works to quantify bias in contextualized word embeddings. We adopt the implementation by Tan and Celis (2019) that uses word-level contextualized embeddings of attribute and stereotype terms instead of sentence-level embeddings. Delobelle et al. (2022) show that using word-level embeddings produces more consistent results and is more robust against the effects of template choices. SEAT is defined as: $s(X_f, X_m, A, B) =$

$\sum_{x_f \in X_f} s(x_f, A, B) - \sum_{x_m \in X_m} s(x_m, A, B)$; A and B are the respective sets of female and male attribute (identity) terms, whereas X_f and X_m are the sets of female and male stereotypes respectively, The similarity measure $s(x, A, B)$ gives the association between a stereotype word's embedding and the word embeddings of attributes: $s(x, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(x, a) - \frac{1}{|B|} \sum_{b \in B} \cos(x, b)$

Log Probability Bias Score (LPBS). This template-based approach proposed by Kurita et al. (2019) quantifies bias using templates containing target words: such as "_ *is a doctor.*". They compute the difference in probability scores that a language model uses to predict words from two respective groups (eg. *female:she, male:he*) to fill in the blank. To account for the effect of prior probabilities of gender words that may skew results, they normalize the results by dividing each prediction by the prior probability of the attribute term. We compute LPBS as: $LPBS = \sum_{x \in X} \sum_i |ls(a_i, x) - ls(b_i, x)|$; where a_i and b_i are equivalent forms of a female and male attribute term respectively, $X = X_f \cup X_m$ is the set of all stereotype words, $ls(w, x) = \log(\frac{P(w|x)}{P(w)})$ is the association between an attribute term $w \in A \cup B$ and a target term $x \in X$, $P(w)$ is the prior probability of w, and $P(w|x)$ is the probability of predicting w as the blank term given the presence of x in its context.

2.3 Extrinsic bias mitigation techniques

We look at two downstream tasks, which are both text classification (Bias-in-Bios (De-Arteaga et al., 2019a) and Jigsaw[6])[7]. To understand the relationship between intrinsic and extrinsic bias, in addition to these original datasets, we include two preprocessed versions of the datasets using *Attribute scrubbing* and *Attribute swapping*. We use these settings to analyze how intrinsic bias is propagated to downstream tasks.

Attribute scrubbing. Attribute features, which are identity terms in an NLP task, are completely removed from text instances in the training data (Prost et al., 2019; De-Arteaga et al., 2019b). This approach may seem intuitive, particularly to mitigate explicit bias, but we suggest that deleting tokens in NLP tasks can be tricky since it could change the syntactic and grammatical structure of the text, leading to out-of-distribution problems.

For datasets such as Bias-in-Bios (De-Arteaga et al., 2019a), a lot of existing works have simply relied on the scrubbed version provided in the dataset, however, we realize that the set of gender words scrubbed is very limited, hence, we extend this to include all gender words from (Kaneko and Bollegala, 2021) and gender names from (Hall Maudslay et al., 2019) to make the approach more effective.

Attribute swapping. CDA has been explored and proposed as another bias mitigation strategy for downstream tasks, particularly in coreference resolution (Lu et al., 2020) and text classification (Park et al., 2018). The idea is to generate counterfactuals by identifying attribute terms in the text instances and swapping them with equivalent terms belonging to the complementary group. We will term this approach *Attribute swapping* to distinguish it from pretrained CDA in Section 2.1. For pretrained CDA, the language model is (further) pretrained on a set of general corpora on Masked Language Modeling task, whereas in attribute swapping, counterfactuals are generated for each instance in the training data based on attribute terms; with the same y labels assigned to both the original and counterfactual examples. Whereas previous versions of CDA ignore names of people, which are major demographic cues, Hall Maudslay et al. (2019) propose a name intervention to generate counterfactuals for

names as well. We use this adaptation for attribute swapping.

2.4 Measuring extrinsic bias

True positive rate difference (TPRD). We use TPRD used in related works (De-Arteaga et al., 2019a; Prost et al., 2019; Jin et al., 2021; Steed et al., 2022). TPRD measures the gap in true positive rates between the predictions for two demographic groups: $TPRD = P(\hat{y} = 1|y = 1, A = a) - P(\hat{y} = 1|y = 1, A = a')$.

Counterfactual Fairness. Since the test dataset follows the same biased distribution of the training set, we also measure the disparity in performance for an individual test instance if the gender attributes are swapped. We measure counterfactual fairness (Kusner et al., 2017) as the difference in the probability of a positive prediction between a test example and its counterfactual, following $CF = \text{mean}(P(|x(\hat{Y}_{A \leftarrow a} = y|A = a, X = x) - P(\hat{Y}_{A \leftarrow a'} = y|A = a, X = x)|))$.

3 Probing for gender information

We design a probe to investigate how much gender information is retained after *mitigating* bias in pretrained language models. Our goal is to measure the amount of gender information in the resulting embeddings of language models after mitigation techniques from § 2.1 have been applied to them.

Defining attribute and stereotype terms. We first identify two sets of attribute terms that define females and males respectively. Let $W_f = \{she, female, woman, ...\}$ and $W_m = \{he, male, man, ...\}$ be the respective sets of female and male gender-defining words. The set $W = W_f \cup W_m$ is the union set of gender-defining words for both females and males. Secondly, we define two sets of stereotype terms which are composed of target words associated with females and males respectively[8]. Let $X_f = \{cheerleader, nurse, softball, ...\}$ and $X_m = \{warrior, baseball, engineer, ...\}$ be the respective sets of female and male stereotype words such that $X = X_f \cup X_m$.

Extracting word embeddings. We extract all sentences from a text corpus containing words in each set of words. Especially for stereotype terms, we ensure that the sentences do not also

---

include words in the attribute wordlist since attribute words in the same context can introduce gender information in the embedding produced. Let $S_w = \{s_{w_1}, ..., s_{w_n}\}$ be the set of all sentences containing a word w. We use the pretrained language model to extract embeddings for w such that $w \rightarrow E_w = \{e_{w_1}, ..., e_{w_n}\}$, where $e_{w_i} \in \mathbb{R}^n$ is the embedding for token w in sentence s_{w_i}. If w appears in s_{w_i} multiple times, then $|E_w| > n$.

Training the classifier. We split W into a train and test sets W_T and W_I respectively such that $W_T \cap W_I = \emptyset$. This is to verify that the trained model is not merely learning to identify words that appear in the train set rather than words containing gender information in general. We reserve the set of stereotypes X solely for evaluating bias (inference) in the language models.

Taking $E_f = \{e_{f1}, e_{f2}, ...\}$ and $E_m = \{e_{m1}, e_{m2}, ...\}$ to be the set of all embeddings for female and male words respectively, given $E = E_f \cup E_m$, we train a classifier $\mathcal{C} : E \rightarrow \{f, m\}$ to predict whether an embedding e belongs to E_f or E_m.

Evaluating bias. To determine the conformity of a language model $\mathcal{L}$ to gender stereotypes X, we introduce two bias criteria (or notions of bias):

Bias accuracy: We compute the accuracy of $\mathcal{C}$ to correctly predict the gender association of all stereotype words in X based on the embeddings produced by $\mathcal{L}$. A high accuracy depicts the presence of pro-stereotype information in the embeddings produced by the language model $\mathcal{L}$, even though the predictions could be of low confidence i.e. just enough gender information to correctly predict the gender association.

Mean bias confidence: We also consider the mean of the individual softmax scores for each prediction. In the case of binary gender, a softmax value of 0.5 means the embedding is perfectly neutral, hence, we define mean bias confidence as $\frac{1}{N} \sum_i^N |bias(e_i) - 0.5|$. High mean confidence depicts an enormous amount of gender information contained in the embedding, although this information may not necessarily be pro-stereotype.

Randomization test. To test the efficacy of our method, We carry out a randomization test by iteratively splitting $X = X_f \cup X_m$ into two random groups $X_{f'}$ and $X_{m'}$. We use $X'^{(i)} = \{X_{f'}^{(i)}, X_{m'}^{(i)}\}$ to define the random pair generated in the ith iteration. We repeat this randomized split over 100

iterations; $X'^{(i)} \in \{X'^{(1)}, ..., X'^{(100)}\}$. By computing a p-value based on a one-sample t-test, we compute if the mean of accuracy scores from the random samples X' is significantly different from the score from $\{X_f, X_m\}$.

4 Investigating intrinsic bias in mitigated models with existing metrics

We first use the two existing metrics defined in § 2.2 (SEAT and LPBS) to quantify bias in BERT-large and ALBERT-large pretrained models. Using the same metrics, we then explore how intrinsic bias changes in mitigated versions of these models using the mitigation techniques outlined in § 2.1: Context-debias, Sent-debias, and CDA pretraining. This is important to evaluate the efficacy of these mitigation strategies, and to make relevant postulations.
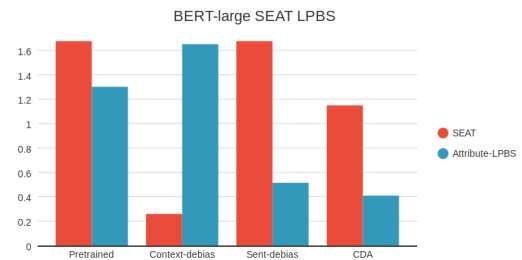
(a) BERT-large model

Figure 2: Intrinsic bias scores using SEAT and LPBS. Results show inconsistencies in measuring bias between SEAT and LPBS for various mitigation strategies — lower scores are desirable in both cases.

4.1 RQ1: Do intrinsic bias measures respond differently to bias mitigation techniques

In Figure 2, we plot the intrinsic bias results[9] of SEAT and LPBS on BERT-large. The first observation we point out, particularly for Context-debias and Sent-debias, is the contradicting nature of bias scores obtained from SEAT and LPBS. Whereas SEAT shows a drastic reduction in intrinsic bias for Context-debias, LPBS indicates worsening bias. The converse is true for Sent-debias where LPBS shows a significant reduction in intrinsic bias but SEAT shows worse scores. This confirms our suspicion that different mitigation strategies respond differently to different metrics.

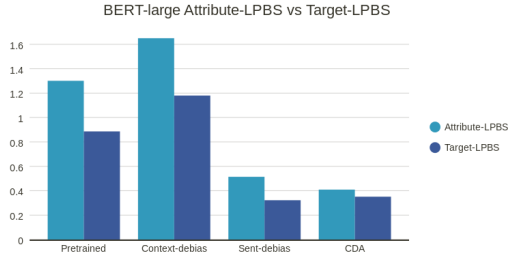[9]see Table 1 and Figure 8 for full results.

Figure 3: We compare LPBS in its default form (attribute-LPBS) and target-LPBS for BERT-large. The plot shows a general positive correlation between the two metrics. In both cases, Sent-debias maintains the lowest bias score — lower scores are desirable in both cases. See Figure 9 for ALBERT-large.

The use of some metrics can be problematic for evaluating some mitigation techniques for intrinsic bias. We find that some metrics are not ideal to use with some mitigation techniques due to differences in how various mitigation techniques interact with a language model. Consider a metric like LPBS which measures the association between a stereotype term and a gender $\frac{P(A=a|X=x)}{P(A=a)}$. This association is captured at the inner layers of the model, but since post-hoc approaches like Sent-debias do not alter the internal representation of the model, LPBS in its original form becomes ineffective — Figure 7. We refer to this original form of LPBS as *attribute-LPBS*. Sent-debias ends up de-biasing the attribute terms rather than the stereotype terms, thus, constantly producing low probability difference scores. Even so, we find LPBS being used in this form with Sent-debias in works such as (Steed et al., 2022). Using stereotype terms instead of attribute terms $P(X=x|A=a)$, which we refer to as *target-LPBS*, is an option to solve this discrepancy, but given that many stereotype terms are usually out of vocabulary, these models will resort to wordpiece tokenization (Sennrich et al., 2016) which will be more challenging to handle. Nonetheless, in Figure 3, when the two versions of LPBS are compared, Sent-debias continues to have the best intrinsic fairness scores in both cases [10].

5 Probing further to uncover bias

We use the methodology in Section 3 to probe for bias in pretrained language models and their bias-

---

[10]For this comparison we only selected single-wordpiece attribute words to avoid challenges with multiple wordpieces.

mitigated versions to see if the intrinsic bias mitigation techniques actually mitigate biases or superficially conceal them from commonplace metrics like SEAT and LPBS. Gonen and Goldberg (2019) discovered that, for non-contextualized word embeddings, the bias mitigation techniques proposed by (Bolukbasi et al., 2016) were hiding bias instead of resolving bias. We use our probe to see if this is equally the case for intrinsic bias mitigation in contextualized embeddings. The consequence of mitigation techniques superficially hiding bias is that downstream classifiers can learn to pick up residual traces of gender information.

We validate our results by carrying out a randomization test, as described in Section 3, to test for statistical significance of our results. We conduct the test under the null hypothesis H_0 that the prediction accuracy of our original gender-stereotype sets will not differ significantly from the random split. We show in Table 2 the very low p-values that indicate statistical significant results, thus refuting H_0.

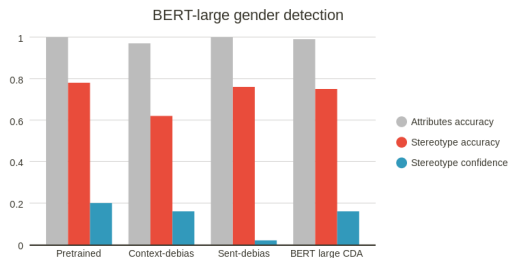### 5.1 RQ2: Intrinsic bias techniques and metrics can hide bias instead of resolving it

Figure 4: Gray bars indicate the accuracy of detecting gender information in attribute terms (high scores are desirable). Red bars indicate the accuracy of detecting gender information in stereotype terms (this should ideally be 0.5 in a fair classifier, showing the inability to correctly predict the gender association of stereotype terms). Blue bars indicate the average confidence of prediction: $\frac{1}{N}\sum_i^N |bias(e_i) - 0.5|$ (low scores are desirable).

In Figure 4, we show the results[11] of our probe for BERT-large. We highlight some key observations.

---

[11]see Table 2 and Figure 10 for full results.

Bias accuracy and mean bias confidence do not correlate. We see from both plots in Figure 4 that, counter-intuitively, high bias accuracy does not concur with high mean bias confidence. Actually, Sent-debias, which consistently has the highest bias accuracy among the mitigation techniques, consistently has the lowest mean bias confidence. This indicates that these two notions of bias may not correspond. This, coupled with the finding in Figure 2 about the inconsistency between SEAT and LPBS, suggests that intrinsic bias could be perceived or measured from conflicting perspectives.

Some mitigation techniques can conceal gender information. For Sent-debias in ALBERT-large, we can, almost to the same degree of accuracy as for the unmitigated pretrained model, correctly predict the association of stereotypes to their respective genders using a simple linear classifier even with a low confidence. This will render a mitigation technique useless if a downstream classifier can correctly predict gender. This could only indicate that although models like Sent-debias significantly reduce the degree of association between stereotypes and their respective genders, hence the low confidence in prediction, this reduction does not inhibit the capability to correctly predict the associated gender of stereotypes; confirming that an intrinsic mitigation process can be rendered useless for downstream fairness even if the degree of association (confidence) is significantly reduced, so long as the information retained remains just enough to correctly associate the gender groups with their associated stereotypes.

5.2 RQ3: Do intrinsic bias mitigation techniques in language models improve fairness in downstream tasks?

To investigate the relationship between intrinsic and extrinsic gender bias, we train both BERT-large and ALBERT-large together with their intrinsic-bias-mitigated versions on the Bias-in-bios and Jigsaw datasets using a 10-fold train-test scheme to correctly predict the profession of each profile. We evaluate extrinsic bias based on gender TPRD and gender CF scores described in § 2.4. We conduct the test under three settings: 1) Using the original version of the training data. 2) Using our attribute scrubbed version Section 2.3. 3) Using the attribute-swap augmented version of the data Section 2.3. These three settings of the dataset are important for us to understand how intrinsic bias

affects downstream fairness under different data settings.

In Figure 5, we show the extrinsic bias score [12] for BERT-large and its *debiased* versions, and the corresponding downstream data type for the Bias-in-Bios dataset — see Figure 11 and Figure 12 for all models and datasets.

Intrinsic mitigation techniques do not show a significant improvement on TPRD. We observe that all the intrinsic bias mitigation interventions we consider do not significantly improve bias in TPRD on their own. There is only a slight improvement or worsening bias in some cases. For Context-debias, we postulate this could be due to the limited list of words used to mitigate bias. If these limited words are not key terms in the downstream task or do not exist in the downstream text instances, the mitigation may not have a consequential effect on the downstream task.

Downstream data processing significantly improves downstream fairness. Secondly, we realize that the downstream data has a lot more material effect on downstream fairness. The data processing techniques: *attribute scrubbing and attribute swapping* showed improvements in TPRD producing the best results.

Combining intrinsic CDA and Downstream CDA produces best TPRD results. Thirdly, we realize that although the downstream data seems key to downstream fairness, a combination of intrinsic and downstream data intervention produces even better extrinsic fairness results.

Sent-debias significantly improves CF scores. Sent-debias is the only intrinsic technique that shows significant improvement in counterfactual fairness. Context-debias and CDA produce worse CF scores in some cases. We observe that the mode of application of Sent-debias in downstream classification tasks does not entirely align with the notion of intrinsic since gender information is removed from the sentence-level representation instead of word-level representation.

Combining Sent-Debias and fair downstream data produces the best CF scores. We also realize from BERT (Figure 5) and ALBERT that combining Sent-debias as a mitigation technique

---

[12] Full results with error margins are given in Table 3, Table 4, Table 5 and Table 6.
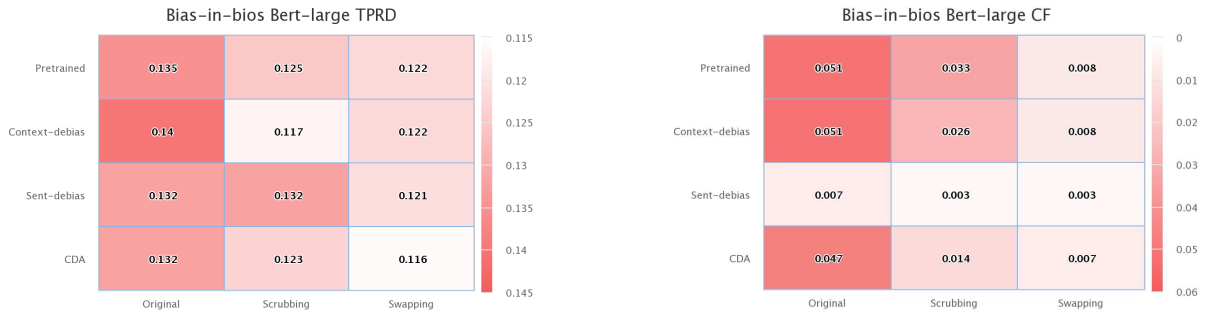
Figure 5: How intrinsic mitigation and downstream data intervention interact to influence fairness on the bias-in-bios data — BERT-large.

and a downstream data intervention produces the best CF scores.
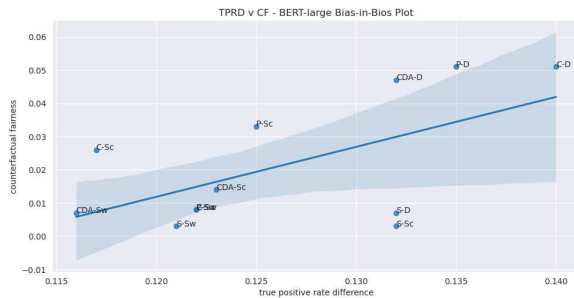

Figure 6: Relationship between TPRD and CF based on results from BERT model and Bias-in-Bios dataset. See Figure 13 for ALBERT-large.

TPRD vs Counterfactual-fairness. Here, we look at the relationship between TPRD and counterfactual fairness. As both measures are used as extrinsic fairness metrics, we aim to see how both results correlate. From Figure 6 we find a positive correlation in general between the two metrics. The downstream data interventions tend to improve both TPRD and CF the most.

6 Related work

Park et al. (2018) study the effects of three techniques to mitigate gender bias in abusive language detection; *debiased word embeddings* (based on the approach by Bolukbasi et al. (2016)), *gender swap* (counterfactual augmentation of classification data), *bias finetuning* using non-contextualized word embeddings: word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). They found that a combination of debiasing and gender swapping produced the best fairness results when implemented with a Gated Recurrent Unit (GRU).

Prost et al. (2019) examine three debiasing techniques for mitigating gender bias in downstream text classification using bias-in-bios (De-Arteaga et al., 2019b), with non-contextualized embeddings. Specifically, they examine *scrubbing* (deletion of identity terms) and approaches that focus on non-contextualized embeddings namely: *debiasing* (based on neutralization and equalization by Bolukbasi et al. (2016)), and *strong debiasing* Prost et al. (2019). Experimenting with Glove embeddings (Pennington et al., 2014), they show that strong debiased produces the best fairness results whilst maintaining a good accuracy score (second only to the unmitigated model). They again show that standard debiasing (Bolukbasi et al., 2016) can be counter-productive in terms of fairness; it can rather reduce fairness in downstream text classification.

Steed et al. (2022) meticulously investigate the effect of intrinsic bias in downstream text classification in RoBERTa. They also conclude that intrinsic/upstream bias does not significantly contribute to downstream bias. However, as we have previously discussed, using the default form of Log probability Bias score (attribute-LPBS) (Kurita et al., 2019) which they adopt to measure intrinsic bias may not be an effective technique to use with Sent-debias.

Orgad et al. (2022), the closest and concurrent with our work, investigate the relationship between intrinsic and extrinsic bias and also develop a probe based on a classification-based technique *minimum description length* (Voita and Titov, 2020) to detect bias in RoBERTa. They reach similar conclusions indicating how intrinsic metrics, like CEAT in their case, could conceal bias.

7 Conclusion

In this paper, we develop a probe to investigate intrinsic bias in two language models, BERT-large and ALBERT-large on two text classification tasks. We use this probe to answer key research questions outlined for this research. We find that intrinsic bias metrics can be sensitive to certain intrinsic bias mitigation techniques. We also show that intrinsic bias mitigation techniques and metrics are capable of concealing bias instead of resolving it. We discover that intrinsic bias mitigation techniques we considered do not significantly improve fairness in text classification when used without other fairness interventions like data pre-processing. We recommend that intrinsic bias mitigation should ideally be combined with other fairness interventions.

Limitations

One obvious limitation of this work is the restriction to the use of English language as a basis. In future, we will explore how these conclusions vary in other language settings. Secondly, although we aimed to make the list of gender and attribute terms used in our experiments as extensive as possible, it is nearly impossible to cover all possible gender and target words in such contexts. Nonetheless, for the scope under consideration, we believe our compilation is extensive enough to give a general outlook and to draw the necessary conclusions. Finally, our work is heavily based on binary notions of gender which is a limitation considering the growing use of non-binary categorizations of gender and other biases which we outline in the ethical considerations section.

Ethical considerations

For practical reasons such as access to datasets and resources on gender bias, we limit our work to a binary representation of gender. We draw readers' attention to the fact that non-binary gender representations are nuanced and intricate (Dev et al., 2021), as such, this should be in cognizance when applying conclusions from this work in non-binary settings. Nonetheless, considering binary gender as a base form of gender categorization, the insights and conclusions from this work can form the baseline for exploring more complex gender categorizations.

References

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

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. *Advances in neural information processing systems*, 29.

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

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019a. Bias in bios: A case study of semantic representation bias in a high-stakes setting. New York, NY, USA. Association for Computing Machinery.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019b. 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*, pages 120–128.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–

1706, Seattle, United States. Association for Computational Linguistics.

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 *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.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

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. In *NAACL*.

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 5267–5275, Hong Kong, China. Association for Computational Linguistics.

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. 2021. Debiasing pre-trained contextualised embeddings. In *Proc. of the 16th European Chapter of the Association for Computational Linguistics (EACL)*.

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.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference*

on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. 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.

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.

Appendix

A   Word lists for experiments

We generate the wordlist we use, given below, from the more extensive wordlist from (Kaneko and Bollegala, 2021). To mitigate the effect of multi-wordpiece tokens, we only select tokens with a single wordpiece. Hence we obtain the following wordlists:

A.1   attribute terms

- **female list:** 'witches', 'mothers', 'diva', 'actress', 'mama', 'dowry', 'princess', 'abbess', 'women', 'widow', 'ladies', 'madam', 'baroness', 'niece', 'lady', 'sister', 'nun', 'her', 'mare', 'convent', 'ladies', 'queen', 'maid', 'chick', 'empress', 'mommy', 'feminism', 'gal', 'estrogen', 'goddess', 'aunt', 'hostess', 'wife', 'mom', 'females', 'ma', 'belle', 'maiden', 'witch', 'miss', 'cow', 'granddaughter', 'her', 'mistress', 'nun', 'actresses', 'girlfriend', 'lady', 'maternal', 'ladies', 'sorority', 'duchess', 'ballerina', 'fiancee', 'wives', 'maternity', 'she', 'heroine', 'queens', 'sisters', 'stepmother', 'daughter', 'lady', 'daughters', 'mistress', 'hostess', 'nuns', 'priestess', 'filly', 'herself', 'girls', 'lady', 'vagina', 'wife', 'mother', 'female', 'womb', 'heiress', 'waitress', 'woman', 'bride', 'grandma', 'bride', 'gal', 'lesbian', 'ladies', 'girl', 'grandmother', 'mare', 'maternity', 'nuns'

- **male list:** 'wizards', 'fathers', 'actor', 'bachelor', 'papa', 'dukes', 'hosts', 'airmen', 'penis', 'prince', 'governors', 'abbot', 'men', 'gentlemen', 'sir', 'baron', 'gods', 'nephew', 'lord', 'brother', 'priest', 'his', 'marquis', 'princes', 'emperors', 'stallion', 'chairman', 'monastery', 'priests', 'king', 'spokesman', 'tailor', 'cowboys', 'dude', 'emperor', 'daddy', 'guys', 'guy', 'godfather', 'god', 'patriarch', 'uncle', 'sir', 'brotherhood', 'host', 'testosterone', 'husband', 'dad', 'steward', 'males', 'pa', 'beau', 'stud', 'bachelor', 'wizard', 'sir', 'bull', 'grandson', 'horsemen', 'rooster', 'bachelor', 'him', 'master', 'lad', 'policeman', 'monk', 'actors', 'boyfriend', 'statesman', 'paternal', 'landlord', 'brethren', 'lords', 'fraternity', 'duke', 'fiance', 'colts', 'husbands', 'he', 'businessman', 'hero', 'deer', 'kings', 'brothers', 'masters', 'stepfather', 'son', 'cowboy', 'sons', 'baritone', 'salesman', 'monks', 'lads', 'congressman', 'priest', 'barons', 'beard', 'colt', 'himself', 'boys', 'lions', 'gentleman', 'penis', 'his', 'bulls', 'uncles', 'lion', 'sorcerer', 'father', 'male', 'sperm', 'prostate', 'businessmen', 'heir', 'waiter', 'headmaster', 'man', 'governor', 'god', 'grandpa', 'groom', 'dude', 'gay', 'boy', 'grandfather', 'priests'

A.2 List of stereotypes(target words)

- **female list:** 'beauty', 'blonde', 'cheerleading', 'cooking', 'flirt', 'flower', 'gossip', 'housekeeper', 'jewelry', 'nanny', 'nurse', 'nurses', 'pink', 'pretty', 'receptionist', 'sewing', 'softball', 'makeup', 'therapist', 'maid', 'chairperson'

- **male list:** 'warrior', 'strong', 'terrorist', 'soldier', 'sniper', 'baseball', 'sergeant', 'bodyguard', 'boss', 'boxer', 'captain', 'carpenter', 'chancellor', 'colonel', 'commander', 'conductor', 'diplomat', 'drummer', 'engineer', 'gangster', 'geek', 'guitarist', 'industrialist', 'marshal', 'mechanic', 'philosopher', 'physicist', 'scientist', 'rapper', 'mechanic', 'carpenter', 'clergy'

B Intrinsic metrics

B.1 Context-debias

$$L_i = \sum_{t \in \mathcal{V}_t} \sum_{x \in \Omega(t)} \sum_{a \in \mathcal{V}_a} (v_i(a)^T E_I(t, x; \theta_e))^2))$$

$$L_{reg} = \sum_{x \in A} \sum_{w \in x} \sum_{i=1} ||E_i(w, x; \theta_e) - E_i(w, x; \theta_p re)||^2$$

$$L = \alpha L_i + \beta L_{reg}$$

Where L_i is the orthogonality constraint such that $E_i(w, x; \theta_e)$ denotes the embedding of token w in the i-th layer of a contextualised word embedding model E with parameters θ_e. $v_i(a)$ is the non-contextualised embedding of an attribute word a. L_{reg} is a regularizer that constrains the Euclidean distance between the contextualized word embedding of a word w in the ith layer in the original pretrained model with parameters θ_{pre}.

B.2 Sent-debias

$$v = PCA_k(\cup_{j=1} \cup_{w \in \mathcal{R}_j} (w - \mu_j))$$

$$h_v = \sum_{j=1}^{k} \langle h, v_j \rangle v_j$$

$$\hat{h} = h - h_v$$

where v represents the top-k gender subspace, h the representation of a given embedding, h_v the projection of h onto v, and $\hat{h}$ the resulting debiased subspace.

C Experiment set-up

We train all models for 4 epochs, a learning rate of $2e^{-5}$ on Tesla V100-SXM3-32GB. We use a sequence length of 100 as the default for all extrinsic tasks. For all downstream tasks, we use a one-layer linear classifier.

D Datasets

Bias-in-bios (De-Arteaga et al., 2019a) consists of online English biographies of people. We use this dataset to predict the occupations of people. Since the gender labels are provided, we can compute disparities in predictive performance between male and female gender groups.

We select the top 7 female and top 7 male professions based on the gender percentages. We take these two sets to represent female and male dominated jobs respectively.

Jigsaw dataset [*https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data*] form the online platform *Civil comments* contains online comments. This dataset is scored from 0 to 10 with the perceived gender of the targets of these comments as well as the toxicity levels of these comments. We use this dataset to predict whether or not a comment is toxic against the target persons. We use gender labels to compute predictive disparities.

We select all comments with toxicity score above 0.5 as the toxic comments and those with a score of exactly 0 as the non-toxic ones in order to eliminate fine-margins. We use the same technique to annotate gender by labeling a gender positive if it scores above 0.5 whilst its complimentary gender has a score of exactly 0.

E Measuring intrinsic bias with LPBS when debiasing is done with Sent-debias

F Results

Table 1: Intrinsic bias scores from SEAT and LPBS. For SEAT we capture the *test statistic score*: **SEAT test score** and the *effect size*: **SEAT eff. size**. For LPBS we capture results from two variants: *attribute-LPBS*: **attr-LPBS** and *target-LPBS*: **target-LPBS**.

(a) Intrinsic bias BERT-large-uncased

Model	SEAT test score	SEAT eff. size	attr-LPBS	target-LPBS
pretrained	0.055	1.675	1.301	0.886
Context-debias	0.002	0.260	1.650	1.180
Sent-debias	0.055	1.675	0.514	0.322
CDA	0.033	1.148	0.409	0.351

(b) Intrinsic bias ALBERT-large

Model	SEAT test score	SEAT eff. size	attr-LPBS	target-LPBS
pretrained	0.051	1.715	0.870	0.446
Context-debias	0.003	0.244	1.343	0.346
Sent-debias	0.051	1.717	0.187	0.111
CDA	0.006	1.170	0.654	0.772

Table 2: Results of gender detection probe: **Gender acc** denotes the accuracy of detecting gender information in attribute/gender terms (high scores are desirable). **Stereotype acc** indicates the accuracy of detecting gender information in stereotype terms (this should ideally be 0.5 in a fair classifier, showing the inability to correctly predict the gender association of stereotype terms). **Stereotype conf** indicate the average confidence of prediction of stereotypes: $\frac{1}{N} \sum_i^N |bias(e_i) - 0.5|$ (low scores are desirable). P-values are from the randomization test.

(a) Detecting gender information BERT-large-uncased

Model	Gender acc	Stereotype acc	Stereotype conf	p-value
pretrained	1.0	0.78	0.20	9.055e-73
Context-debias	0.97	0.62	0.16	1.398e-51
Sent-debias	1.0	0.76	0.02	8.772e-70
CDA	0.99	0.75	0.16	7.824e-77

(b) Detecting gender information Alert-large-uncased

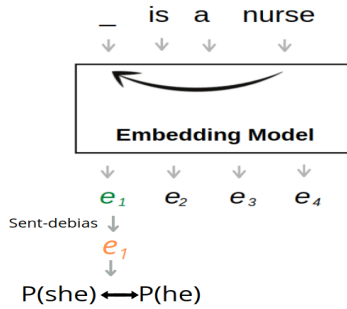| Model	Gender acc	Stereotype acc	Stereotype conf	p-value
pretrained	1.0	0.83	0.33	2.763e-71
Context-debias	0.97	0.63	0.19	1.700e-48
Sent-debias	1.0	0.82	0.02	2.335e-95
CDA	0.94	0.63	0.12	3.497e-61

Figure 7: Measuring intrinsic bias in Sent-debias with target-LPBS.

Table 3: Extrinsic bias scores from Bias in Bios - BERT-large: We capture true positive rate difference, false positive rate difference, counterfactual fairness, accuracy scroes for both gender groups, counterfactual augmented true positive rate difference and counterfactual augmented false positive rate diference scores.

Model	Data	TPRD	FPRD	ACC-F	ACC-M	CF	CF-TPRD	CF-FPRD	CF-ACC-F	CF-ACC-M
Pretrained	Default	0.135± 0.03	0.012± 0.0	0.984± 0.0	0.982± 0.0	0.051± 0.0	0.116± 0.02	0.008± 0.0	0.982± 0.0	0.981± 0.0
Pretrained	Scrubbing	0.125 ± 0.02	0.011 ± 0.0	0.984 ± 0.0	0.982 ± 0.0	0.033 ± 0.0	0.113 ± 0.02	0.008 ± 0.0	0.983 ± 0.0	0.981 ± 0.0
Pretrained	Swapping	0.122 ± 0.02	0.008 ± 0.0	0.983 ± 0.0	0.981 ± 0.0	0.008 ± 0.0	0.12 ± 0.02	0.008 ± 0.0	0.983 ± 0.0	0.981 ± 0.0
Context-debias	Default	0.14 ± 0.03	0.012 ± 0.0	0.984 ± 0.0	0.982 ± 0.0	0.051 ± 0.0	0.119 ± 0.02	0.008 ± 0.0	0.982 ± 0.0	0.98 ± 0.0
Context-debias	Scrubbing	0.117 ± 0.03	0.01 ± 0.0	0.984 ± 0.0	0.981 ± 0.0	0.026 ± 0.0	0.109 ± 0.03	0.009 ± 0.0	0.983 ± 0.0	0.981 ± 0.0
Context-debias	Swapping	0.122 ± 0.03	0.009 ± 0.0	0.983 ± 0.0	0.981 ± 0.0	0.008 ± 0.0	0.12 ± 0.03	0.009 ± 0.0	0.983 ± 0.0	0.981 ± 0.0
CDA	Default	0.132 ± 0.02	0.012 ± 0.0	0.983 ± 0.0	0.982 ± 0.0	0.047 ± 0.0	0.111 ± 0.02	0.008 ± 0.0	0.982 ± 0.0	0.98 ± 0.0
CDA	Scrubbing	0.123 ± 0.03	0.009 ± 0.0	0.982 ± 0.0	0.981 ± 0.0	0.014 ± 0.0	0.121 ± 0.03	0.009 ± 0.0	0.982 ± 0.0	0.981 ± 0.0
CDA	Swapping	0.116 ± 0.02	0.009 ± 0.0	0.983 ± 0.0	0.981 ± 0.0	0.007 ± 0.0	0.116 ± 0.02	0.009 ± 0.0	0.983 ± 0.0	0.981 ± 0.0
Sent-debias	Original	0.132 ± 0.02	0.012 ± 0.0	0.983 ± 0.0	0.982 ± 0.0	0.007 ± 0.0	0.115 ± 0.02	0.008 ± 0.0	0.981 ± 0.0	0.98 ± 0.0
Sent-debias	Scrubbing	0.132 ± 0.02	0.01 ± 0.0	0.983 ± 0.0	0.981 ± 0.0	0.003 ± 0.0	0.122 ± 0.03	0.008 ± 0.0	0.982 ± 0.0	0.98 ± 0.0
Sent-debias	Swapping	0.121 ± 0.02	0.009 ± 0.0	0.982 ± 0.0	0.98 ± 0.0	0.003 ± 0.0	0.119 ± 0.02	0.009 ± 0.0	0.982 ± 0.0	0.98 ± 0.0

Table 4: Extrinsic bias scores from Bias in Bios - ALBERT-large

Model	Data	TPRD	FPRD	ACC-F	ACC-M	CF	CF-TPRD	CF-FPRD	CF-ACC-F	CF-ACC-M
Pretrained	Default	0.138 ± 0.02	0.013 ± 0.0	0.982 ± 0.0	0.979 ± 0.0	0.06 ± 0.0	0.118 ± 0.02	0.009 ± 0.0	0.979 ± 0.0	0.978 ± 0.0
Pretrained	Scrubbing	0.127 ± 0.01	0.01 ± 0.0	0.982 ± 0.0	0.978 ± 0.0	0.021 ± 0.0	0.12 ± 0.01	0.009 ± 0.0	0.981 ± 0.0	0.978 ± 0.0
Pretrained	Swapping	0.121 ± 0.02	0.009 ± 0.0	0.981 ± 0.0	0.979 ± 0.0	0.01 ± 0.0	0.121 ± 0.02	0.009 ± 0.0	0.981 ± 0.0	0.979 ± 0.0
Context-debias	Default	0.137 ± 0.02	0.012 ± 0.0	0.983 ± 0.0	0.98 ± 0.0	0.062 ± 0.0	0.114 ± 0.02	0.009 ± 0.0	0.981 ± 0.0	0.978 ± 0.0
Context-debias	Scrubbing	0.129 ± 0.02	0.01 ± 0.0	0.98 ± 0.0	0.978 ± 0.0	0.021 ± 0.0	0.124 ± 0.02	0.009 ± 0.0	0.979 ± 0.0	0.977 ± 0.0
Context-debias	Swapping	0.116 ± 0.02	0.009 ± 0.0	nan ± nan	nan ± nan	0.011 ± 0.0	0.115 ± 0.02	0.009 ± 0.0	0.981 ± 0.0	0.978 ± 0.0
CDA	Default	0.142 ± 0.02	0.013 ± 0.0	0.981 ± 0.0	0.978 ± 0.0	0.063 ± 0.0	0.121 ± 0.02	0.009 ± 0.0	0.979 ± 0.0	0.977 ± 0.0
CDA	Scrubbing	0.127 ± 0.02	0.01 ± 0.0	0.979 ± 0.0	0.977 ± 0.0	0.012 ± 0.0	0.125 ± 0.02	0.01 ± 0.0	0.979 ± 0.0	0.977 ± 0.0
CDA	Swapping	0.115 ± 0.01	0.009 ± 0.0	0.98 ± 0.0	0.977 ± 0.0	0.011 ± 0.0	0.116 ± 0.02	0.009 ± 0.0	0.98 ± 0.0	0.977 ± 0.0
Sent-debias	Original	0.145 ± 0.02	0.013 ± 0.0	0.981 ± 0.0	0.978 ± 0.0	0.007 ± 0.0	0.119 ± 0.03	0.009 ± 0.0	0.978 ± 0.0	0.976 ± 0.0
Sent-debias	Scrubbing	0.12 ± 0.02	0.01 ± 0.0	0.978 ± 0.0	0.976 ± 0.0	0.002 ± 0.0	0.108 ± 0.02	0.009 ± 0.0	0.978 ± 0.0	0.976 ± 0.0
Sent-debias	Swapping	0.117 ± 0.03	0.009 ± 0.0	0.981 ± 0.0	0.978 ± 0.0	0.002 ± 0.0	0.114 ± 0.02	0.009 ± 0.0	0.981 ± 0.0	0.978 ± 0.0

Table 5: Extrinsic bias scores from Jigsaw - BERT-large

Model	Data	TPRD	FPRD	ACC-F	ACC-M	CF	CF-TPRD	CF-FPRD	CF-ACC-F	CF-ACC-M
Pretrained	Default	0.044 ± 0.02	0.118 ± 0.04	0.896 ± 0.01	0.934 ± 0.0	0.116 ± 0.01	0.061 ± 0.01	0.029 ± 0.02	0.866 ± 0.01	0.902 ± 0.01
Pretrained	Scrubbing	0.026 ± 0.02	0.076 ± 0.03	0.891 ± 0.01	0.914 ± 0.01	0.031 ± 0.0	0.026 ± 0.01	0.058 ± 0.02	0.889 ± 0.01	0.904 ± 0.01
Pretrained	Swapping	0.027 ± 0.01	0.065 ± 0.02	0.892 ± 0.01	0.92 ± 0.01	0.01 ± 0.0	0.029 ± 0.01	0.066 ± 0.01	0.893 ± 0.01	0.921 ± 0.01
Context-debias	Default	0.047 ± 0.02	0.128 ± 0.04	0.895 ± 0.01	0.931 ± 0.01	0.116 ± 0.01	0.049 ± 0.02	0.032 ± 0.02	0.868 ± 0.01	0.894 ± 0.01
Context-debias	Scrubbing	0.014 ± 0.01	0.032 ± 0.02	0.701 ± 0.22	0.874 ± 0.06	0.015 ± 0.01	0.015 ± 0.01	0.026 ± 0.02	0.698 ± 0.21	0.87 ± 0.06
Context-debias	Swapping	0.031 ± 0.01	0.049 ± 0.02	0.889 ± 0.01	0.915 ± 0.01	0.01 ± 0.0	0.031 ± 0.01	0.05 ± 0.02	0.888 ± 0.01	0.915 ± 0.01
CDA	Default	0.055 ± 0.02	0.114 ± 0.03	0.902 ± 0.01	0.932 ± 0.0	0.115 ± 0.01	0.048 ± 0.01	0.024 ± 0.01	0.872 ± 0.01	0.899 ± 0.0
CDA	Scrubbing	0.023 ± 0.01	0.053 ± 0.03	0.894 ± 0.01	0.904 ± 0.01	0.018 ± 0.0	0.022 ± 0.01	0.041 ± 0.02	0.892 ± 0.01	0.899 ± 0.01
CDA	Swapping	0.026 ± 0.02	0.038 ± 0.02	0.891 ± 0.01	0.913 ± 0.01	0.008 ± 0.0	0.026 ± 0.02	0.036 ± 0.02	0.891 ± 0.01	0.913 ± 0.01
Sent-debias	Original	0.074 ± 0.02	0.108 ± 0.03	0.898 ± 0.01	0.926 ± 0.01	0.069 ± 0.01	0.041 ± 0.02	0.025 ± 0.01	0.865 ± 0.01	0.89 ± 0.01
Sent-debias	Scrubbing	0.033 ± 0.01	0.035 ± 0.01	0.889 ± 0.01	0.909 ± 0.01	0.012 ± 0.0	0.037 ± 0.01	0.023 ± 0.02	0.885 ± 0.01	0.901 ± 0.01
Sent-debias	Swapping	0.021 ± 0.02	0.045 ± 0.01	0.893 ± 0.01	0.916 ± 0.01	0.006 ± 0.0	0.021 ± 0.02	0.045 ± 0.01	0.893 ± 0.01	0.915 ± 0.01

Table 6: Extrinsic bias scores from Jigsaw - ALBERT-large

Model	Data	TPRD	FPRD	ACC-F	ACC-M	CF	CF-TPRD	CF-FPRD	CF-ACC-F	CF-ACC-M
Pretrained	Default	0.077 ± 0.04	0.115 ± 0.04	0.894 ± 0.01	0.916 ± 0.01	0.143 ± 0.01	0.035 ± 0.02	0.039 ± 0.02	0.862 ± 0.01	0.86 ± 0.02
Pretrained	Scrubbing	0.032 ± 0.02	0.025 ± 0.02	0.789 ± 0.14	0.819 ± 0.1	0.031 ± 0.01	0.035 ± 0.03	0.033 ± 0.02	0.783 ± 0.14	0.808 ± 0.1
Pretrained	Swapping	0.037 ± 0.02	0.029 ± 0.01	0.869 ± 0.02	0.895 ± 0.02	0.012 ± 0.0	0.038 ± 0.02	0.032 ± 0.02	0.869 ± 0.02	0.895 ± 0.02
Context-debias	Default	0.064 ± 0.02	0.122 ± 0.02	0.891 ± 0.01	0.92 ± 0.01	0.137 ± 0.01	0.051 ± 0.02	0.025 ± 0.01	0.857 ± 0.01	0.875 ± 0.01
Context-debias	Scrubbing	0.031 ± 0.01	0.05 ± 0.03	0.885 ± 0.01	0.891 ± 0.01	0.039 ± 0.01	0.024 ± 0.01	0.041 ± 0.02	0.878 ± 0.01	0.874 ± 0.01
Context-debias	Swapping	0.023 ± 0.01	0.048 ± 0.02	0.874 ± 0.01	0.904 ± 0.01	0.011 ± 0.0	0.025 ± 0.01	0.048 ± 0.02	0.875 ± 0.01	0.905 ± 0.01
CDA	Default	0.078 ± 0.02	0.115 ± 0.03	0.887 ± 0.01	0.913 ± 0.01	0.139 ± 0.01	0.043 ± 0.02	0.031 ± 0.01	0.855 ± 0.01	0.865 ± 0.01
CDA	Scrubbing	0.026 ± 0.01	0.052 ± 0.02	0.866 ± 0.02	0.865 ± 0.02	0.024 ± 0.0	0.021 ± 0.01	0.036 ± 0.02	0.867 ± 0.02	0.857 ± 0.02
CDA	Swapping	0.042 ± 0.02	0.038 ± 0.02	0.866 ± 0.01	0.9 ± 0.01	0.011 ± 0.0	0.041 ± 0.02	0.037 ± 0.02	0.867 ± 0.01	0.9 ± 0.01
Sent-debias	Original	0.102 ± 0.05	0.106 ± 0.05	0.765 ± 0.18	0.87 ± 0.05	0.064 ± 0.03	0.023 ± 0.02	0.03 ± 0.02	0.736 ± 0.17	0.824 ± 0.03
Sent-debias	Scrubbing	0.011 ± 0.01	0.021 ± 0.02	0.685 ± 0.21	0.831 ± 0.04	0.009 ± 0.01	0.011 ± 0.01	0.018 ± 0.02	0.683 ± 0.21	0.825 ± 0.03
Sent-debias	Swapping	0.031 ± 0.02	0.031 ± 0.02	0.821 ± 0.13	0.88 ± 0.04	0.006 ± 0.0	0.034 ± 0.02	0.031 ± 0.02	0.818 ± 0.13	0.88 ± 0.04

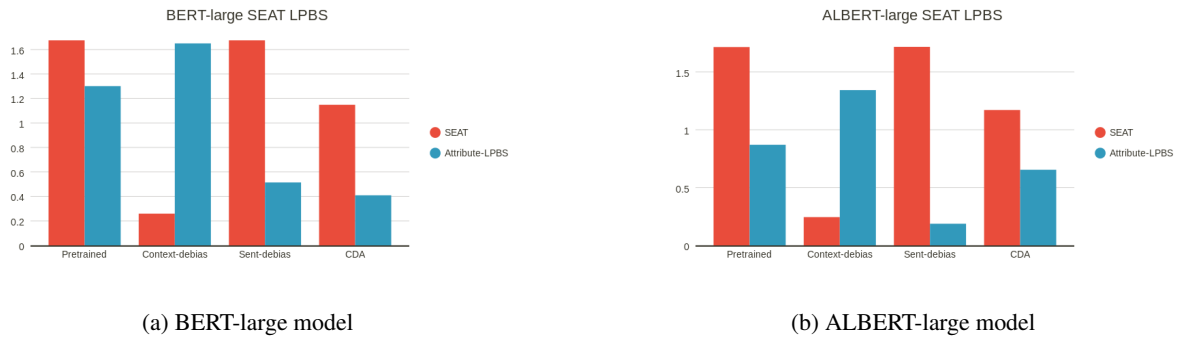(a) BERT-large model

(b) ALBERT-large model

Figure 8: Intrinsic bias scores using SEAT and LPBS. Results show inconsistencies in measuring bias between SEAT and LPBS for various mitigation strategies — lower scores are desirable in both cases.
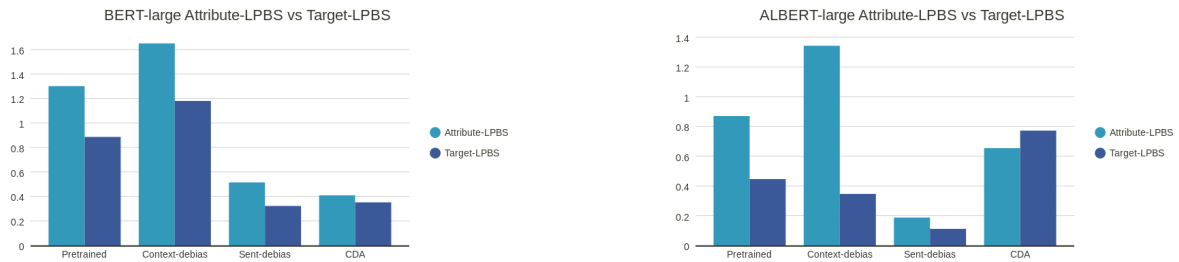

Figure 9: We compare LPBS in its default form (attribute-LPBS) and target-LPBS for both BERT and ALBERT. The plot shows a general positive correlation between the two metrics. In both cases, Sent-debias maintains the lowest bias score — lower scores are desirable in both cases
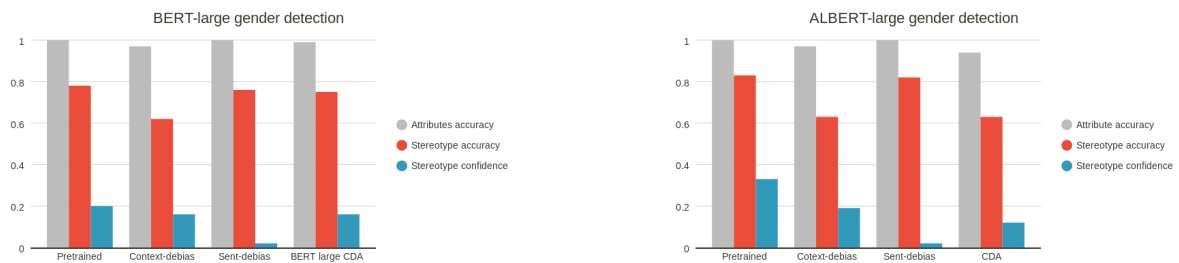

Figure 10: Gray bars indicate the accuracy of detecting gender information in attribute terms (high scores are desirable). Red bars indicate the accuracy of detecting gender information in stereotype terms (this should ideally be 0.5 in a fair classifier, showing the inability to correctly predict the gender association of stereotype terms). Blue bars indicate the average confidence of prediction: $\frac{1}{N} \sum_{i}^{N} |bias(e_i) - 0.5|$ (low scores are desirable).

Figure 11: How intrinsic bias mitigation and downstream data intervention interact to influence fairness on the bias-in-bios dataset.

Figure 12: How intrinsic mitigation and downstream data intervention interact to influence counterfactual fairness on the jigsaw data.
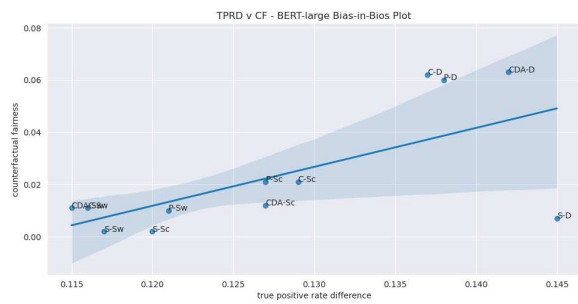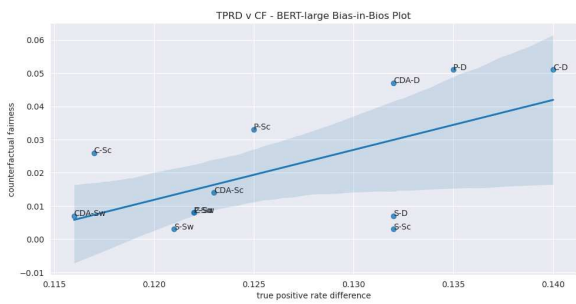
Figure 13: Relationship between TPRD and CF based on results from the Bias-in-Bios dataset.