Don't Just Clean It, Proxy Clean It: Mitigating Bias by Proxy in Pre-Trained Models

Swetasudha Panda and **Ari Kobren** and **Michael Wick** and **Qinlan Shen**
Oracle Labs
Burlington, MA
{swetasudha.panda, ari.kobren, michael.wick, qinlan.shen}@oracle.com

Abstract

Transformer-based pre-trained models are known to encode societal biases, not only in their contextual representations but also in their downstream predictions when fine-tuned on task-specific data. We present D-BIAS, an approach that selectively eliminates stereotypical associations (e.g, co-occurrence statistics) at fine-tuning, such that the model doesn't learn to excessively rely on those signals. D-BIAS attenuates biases from both identity words *and* frequently co-occurring proxies, which we select using pointwise mutual information. We apply D-BIAS to a) occupation classification, and b) toxicity classification and find that our approach substantially reduces downstream biases ($> 60\%$ in toxicity classification for identities that are most frequently flagged as toxic on online platforms). In addition, we show that D-BIAS dramatically improves upon *scrubbing*, i.e., removing only the identity words in question. We also demonstrate that D-BIAS easily extends to multiple identities and achieves competitive performance with two recently proposed debiasing approaches: R-LACE and INLP.

1 Introduction

One of the most dominant paradigms in natural language processing is the fine-tuning of large language models for downstream tasks (Bommasani et al., 2021). Transformer-based language models pre-trained on massive volumes of text achieve state-of-the-art performance on a wide spectrum of tasks, when fine-tuned on task-specific data (Radford et al., 2018; Devlin et al., 2019). A major concern of this paradigm, however, is that pre-trained models can learn societal biases from both the pre-training and fine-tuning data. These biases can be expressed in two ways – in a) the contextualized representations of the model themselves, i.e., *intrinsic biases* (Nangia et al., 2020; Nadeem et al., 2021; May et al., 2019; Kurita et al., 2019), and b) its downstream predictions after fine-tuning, i.e., *extrinsic* or *allocational harms* (Gehman et al., 2020; Garimella et al., 2019; Blodgett et al., 2018).

There has been extensive research into mitigating bias in pre-trained language models (Meade et al., 2022). Prior work has primarily focused on debiasing contextual representations upstream, i.e., before the pre-trained model is fine-tuned (Liang et al., 2020; Kaneko and Bollegala, 2021; Ravfogel et al., 2020, 2022). However, recent work on the bias-transfer hypothesis[1] (Steed et al., 2022) reveals that most of the variation in downstream biases can be explained by biased associations in the fine-tuning dataset. This work also shows that variations in upstream bias have little impact on downstream disparities. These results emphasize the importance of debiasing interventions that target biases[2] introduced during fine-tuning from task-specific data.

However, Steed et al. (2022) also show that simply scrubbing identity terms [3] from the downstream/ fine-tuning data does not work for bias mitigation. Scrubbing the fine-tuning data is effective in reducing bias only when the model is not pre-trained (i.e., a model with randomly initialized weights). This finding emphasizes that while an intervention targeted at the fine-tuning data is crucial, pre-training can complicate downstream-focused debiasing interventions.

We hypothesize that the ineffectiveness of scrubbing the fine-tuning data on a pre-trained model is because pre-trained models are more sensitive to the *bias by proxy* problem, at fine-tuning. This problem arises when words that co-occur with an identity term function as proxies for that identity term and therefore, act as an additional source of bias (De-Arteaga et al., 2019). Typically, a pre-

[1]The hypothesis that social biases internalized by large language models during pre-training transfer into harmful task-specific behavior after fine-tuning

[2]Referring to statistical associations for social stereotypes

[3]Words that indicate demographic identities

trained model predicts masked tokens or spans from their contexts. During pre-training, sometimes an identity term is masked and the model learns to predict it from any proxy words that happen to appear in the identity's context. Similarly, sometimes an identity word is part of the context for predicting a masked proxy word. In this way, we suspect that the pre-training process entangles identity terms with words that could serve as a proxy at fine-tuning. In contrast, a model that is not pre-trained has randomly initialized weights, and therefore, there is no initial entangling of identity words with their proxies. This can explain the effectiveness of simply scrubbing the identity terms, in the absence of pre-training.

Based on the bias by proxy intuition, to address downstream biases, we present Dropout BIas ASsociations (D-BIAS), an approach which selectively *drops out* or ignores bias-associated words and proxies at fine-tuning, in addition to scrubbing identity words. Specifically, D-BIAS aims to identify words, in the vocabulary of the fine-tuning data, that are most likely to act as proxies for identity words in the context of a particular downstream task. We identify such proxies based on co-occurrence frequencies using pointwise mutual information (Church and Hanks, 1990). During fine-tuning, D-BIAS replaces identity words and a set of the most relevant proxies with `[MASK]` (details in section 2), preventing the model from relying on bias associations *and* also preventing it from inferring those associations from the proxies. As a result, our approach is simple and relatively straightforward to implement. Moreover, it can readily and effectively be extended to multiple identity groups.

We experiment with D-BIAS and other variations of word dropout on two downstream tasks: (multi-class) occupation classification of online biographies (De-Arteaga et al., 2019) and toxicity classification on Wikipedia Talk Page comments (Dixon et al., 2018). We demonstrate that our debiasing approaches are consistently effective in reducing downstream biases compared to fine-tuning without word dropout. Improvements are especially dramatic with D-BIAS on toxicity classification. For certain identity groups that are more vulnerable to unfair censorship on online forums, such as `gay` and `homosexual`, D-BIAS reduces classifier false positive rates (FPR) by $> 60\%$ and even improves classification accuracy in the process. In contrast, scrubbing these identity terms does not decrease the FPR for these groups. In occupation classification, all of our proposed methods outperform two state-of-the-art debiasing approaches R-LACE (Ravfogel et al., 2022) and INLP (Ravfogel et al., 2020), in terms of reducing the true positive rate (TPR) gaps between biographies of men and women across occupations.

2 Methods

Our hypothesis is centered on addressing proxy behavior in the downstream data, in order to prevent the model from learning or inferring societal stereotypes. For a given identity word, we define proxies as a set of words which frequently co-occur with the identity word, in the fine-tuning vocabulary. We describe our approach to compute these proxy words more formally in section 2.1.

Our proposed approach is straightforward and aims to find sets of the most relevant proxies corresponding to identity words. Accordingly, it consists of three major components. The first step is to analyze the fine-tuning data to identify proxy words that are likely to be associated with identity terms. We describe our approach for identifying proxies in section 2.1. The second step is to employ word dropout strategies, building on those word associations. More precisely, we define word dropout as replacing a word with the `[MASK]` token.[4] Our word dropout strategies are based on specific heuristics that utilize information on proxies, to make decisions on whether a word in the fine-tuning vocabulary should be dropped out. We present these strategies in section 2.2. The third step is the actual fine-tuning of the pre-trained model, on a version of the fine-tuning data in which identity words and relevant proxies have been dropped out (e.g., using one of the dropout strategies).

We begin with a description of the downstream tasks, in order to refer to concrete examples.

Occupation Classification In this task, the goal is to identify a person's occupation from their online biography. We consider the bias-in-bios (BIOS) data (De-Arteaga et al., 2019), which consists of $\sim$ 400K online biographies for 28 occupations, scraped from Common Crawl. De-Arteaga et al. (2019) and Steed et al. (2022) report that models trained on this data exhibit disparities in the em-

[4] We found that there is no significant difference when replacing with '_' and that `[MASK]` works better compared to an `[OOV]` token.

pirical true positive rates (TPR) within an occupation, for biographies belonging to men and women. TPR denotes the likelihood that the classifier correctly identifies a person's occupation from their biography. More specifically, biographies with `she/her` pronouns are less frequently classified as male-dominated professions—such as `surgeon` (and vice versa for occupations such as `model`). These discrepancies in the model predictions can lead to allocational harms — e.g., in recruitment scenarios. We follow prior work and consider binary gender as the bias-related identity. We refer to the pronouns used in each biography (each biography uses either `he/him` or `she/her` pronouns) as the identity words, following De-Arteaga et al. (2019).

Toxicity Classification The goal of this task is to identify toxic comments posted on an online forum. We consider the WIKI dataset which consists of $\sim$ 128K comments from Wikipedia Talk Pages (Dixon et al., 2018). Each comment is labeled by human raters as `Toxic` or `Not Toxic`.[5] Dixon et al. (2018) outline $\sim$ 50 common demographic identities (Table 4) based on gender, age, ethnicity, disability status and religion, many of which have a very skewed representation in the data (Figure 7). For example, the identity term `gay` appears in only 0.5% of all comments, but it appears in 3% of the comments labeled as `Toxic`. Because of the disproportionate number of toxic examples for comments containing identity words such as `gay`, `queer` or `homosexual`, these mentions are more likely to be flagged for toxic content, resulting in discriminatory censorship for comments that mention these groups. We follow Dixon et al. (2018); Steed et al. (2022) and consider the empirical false positive rate (FPR) as a measure of the model's likelihood to falsely flag a neutral comment as `Toxic`.

2.1 Identifying Proxies with Mutual Information

Let A denote the set of identity words in the training set of the fine-tuning data (we will refer to this as the training data). In BIOS, A is the set of pronouns in De-Arteaga et al. (2019) and in WIKI, A is a set of identity words outlined in Dixon et al. (2018) (Table 4). To find proxies corresponding to the identity groups, we propose to compute the point-wise mutual information (PMI) (Church and Hanks, 1990) between the identity words and words in the vocabulary of the training data.

The PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the probability of their coincidence, given their joint distribution and their individual distributions, assuming independence:

$$\text{PMI}(x;y) = \log\left[p(x,y)/p(x)p(y)\right]$$

For a pair of words, PMI quantifies the likelihood of their co-occurrence, taking into account the probability of single occurrences of each. A high PMI score between two words indicates a high probability of the words co-occurring together and a lower probability of either one of those (or both) occurring singularly. We normalize PMIs to be $\in [-1,+1]$, where -1 (in the limit) indicates no co-occurrence (i.e., for words which never co-occur), 0 indicates independence (i.e., equal chance of co-occurring or not), and $+1$ indicates complete co-occurrence (i.e., for words that always co-occur):

$$\text{nPMI}(x;y) = \text{PMI}(x;y)/h(x,y)$$

where $h(x,y)$ is the joint self-information, estimated as $-\log p(X=x,Y=y)$.

To compute a set of proxies for each identity word, we consider each word w in the vocabulary of the fine-tuning data, and compute a set of nPMI scores $\text{nPMI}(w;a)$ with each of the identity words $a \in A$. In case of the BIOS data, we group all `he/his` set of pronouns and `she/her` set of pronouns to compute the co-occurrences and the nPMI collectively for each group of pronouns.

2.2 Word Dropout Strategies

In this section, based on the nPMI computations, we outline four heuristics for making word dropout decisions. In the first two approaches, we consider all words in the vocabulary of the fine-tuning data as candidates for dropout, as long as these words co-occur at least once with any of the identity words. In the next two approaches, we consider word dropout decisions at the document level, where a document refers to an instance of the fine-tuning data.

D-BIAS: For each word w in the training data vocabulary, D-BIAS makes dropout decisions based on its nPMI with each identity word in A. D-BIAS

[5] The authors define a toxic comment as a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion".

drops out w if the max of the nPMI scores over all identity words, i.e., $\max_{a \in A} \text{nPMI}(w; a) \geq \theta$, where θ is a hyperparameter which we select using the validation data at fine-tuning.

D-BIAST: This approach is a stochastic variation of D-BIAS which makes dropout decision for each word w based on a probability proportional to the maximum over nPMIs with all identity words. Let $s(w) = \max_{a \in A} \text{nPMI}(w; a)$. Next, let $\min_{\texttt{nPMI}}$ and $\max_{\texttt{nPMI}}$ be the minimum and maximum of the nPMI scores over the fine-tuning vocabulary. More formally,

$$\min_{\texttt{nPMI}} = \min_{w \in W, a \in A} \text{nPMI}(w; a),$$
$$\max_{\texttt{nPMI}} = \max_{w \in W, a \in A} \text{nPMI}(w; a).$$

The dropout probability $p(w)$ for each word w is then computed as:

$$p(w) = \big(s(w) - \min_{\texttt{nPMI}}\big) / \big(\max_{\texttt{nPMI}} - \min_{\texttt{nPMI}}\big).$$

Next, we propose two approaches for word dropout at the document level, where a document is an individual biography in case of BIOS and a Talk Page comment in WIKI. In these approaches, a word in a document is considered for dropout if it has a high nPMI with an identity word mentioned in that document. For example, in WIKI task, if an online forum comment mentions `gay`, we dropout words that have high nPMIs with `gay`, within that comment. The previous two approaches, on the other hand, dropout all mentions of a word in the fine-tuning vocabulary, if it has a high nPMI with any of the 50 identity words.

SENT-K: This approach selects the set of k highest nPMI words (with repetition, within a document of the training data), for each identity word mentioned in that document. Here k is a hyperparameter which we select using the validation set.

SENT-ST: This approach is a stochastic variation of SENT-K. For any mention of identity word $a \in A$ within a document in the training set, it drops out words w in that document with a probability:

$$p(w) = \big(\text{nPMI}(w; a) - \min_{\texttt{nPMI}}\big) / \big(\max_{\texttt{nPMI}} - \min_{\texttt{nPMI}}\big)$$

where $\min_{\texttt{nPMI}}$ and $\max_{\texttt{nPMI}}$ denote the range of nPMI scores computed over each word w in that sentence (with the identity word a).

In each case, we fine-tune the model on a *masked* dataset where the identity tokens and the proxies have been replaced with `[MASK]` using one of the dropout strategies.

3 Experiments

3.1 Setup

We experiment with BERT-base-uncased (Devlin et al., 2019). In each experiment i.e., with or without dropout strategies, we fine-tune BERT for 5 epochs and select the best model based on validation accuracy.[6] For D-BIAS and SENT-K, we choose hyperparameters θ and k based on accuracy on the validation set. Recall that for D-BIAS, θ is a threshold on the nPMI scores, for a word to be considered for dropout. For SENT-K, k represents the top k highest nPMI words to dropout within a document of the fine-tuning data, corresponding to any identity words in that document. (section 2.2). We experiment with $\theta \in (0.0, 0.7)$ for WIKI and $\theta \in (0.01, 0.9)$ for BIOS (based on the range of the nPMIs in each case), and we consider $k = 5, 10$ and 20 for SENT-K.

Additionally, we test two variations of nPMI computations. In the first, we compute co-occurrences within each sentence (using a sentence tokenizer[7]). In the second variation, we compute co-occurrences within each training document (e.g., an instance of a biography has 4 sentences on an average). In each case, we report best results based on performance on the validation set. We omit stop words with NLTK and also punctuation from being considered for dropout (Bird et al., 2009).

We compare our approaches with two state-of-the-art debiasing approaches: R-LACE and INLP (Ravfogel et al., 2022, 2020). INLP iteratively identifies a linear subspace corresponding to biases (gender biases) and subtracts projections in the embedding space. R-LACE formulates this problem of identifying and subtracting a linear subspace such that a linear predictor can not recover the subtracted subspace. More precisely, it formulates the problem as a constrained, linear minimax game, and derives a closed-form solution. R-LACE outperforms INLP in finding a minimal rank bias subspace.

In addition to the above approaches, we compare results with the following two baselines:

[6] On WIKI data, we find similar results with 10, 20 and 30 training epochs

[7] `nltk.sent_tokenize()`

	Test Acc ↑	TPR$_{\text{gap}}$(RMSE) ↓	ρ TPR$_{\text{gap}}$%$_{\mathbf{F}}$ ↓
BERT	86.04 (0.10)	0.145 (0.005)	0.818 (0.005)
D-BIAS	84.40 (0.25)	**0.088 (0.006)**	0.728 (0.029)
D-BIAST	84.67 (0.24)	0.112 (0.003)	0.738 (0.024)
SENT-K	**85.93 (0.06)**	0.105 (0.004)	**0.719 (0.014)**
SENT-ST	85.90 (0.10)	0.101 (0.003)	**0.719 (0.022)**
UNIFORM	85.36 (0.19)	0.110 (0.006)	0.741 (0.017)
SCRUB	85.90 (0.04)	0.103 (0.003)	0.720 (0.021)
INLP	84.98 (0.06)	0.113 (0.009)	0.797 (0.027)
R-LACE:1	85.09 (0.07)	0.117 (0.011)	0.794 (0.025)
R-LACE:100	85.04 (0.09)	0.115 (0.014)	0.792 (0.025)

Table 1: **Debiasing Results on BIOS** means and variances (in parenthesis) over 5 random initializations. Test Acc is overall accuracy on the testset (higher is better). TPR gap is the RMSE of the TPR gap across occupations (lower is better). ρ is correlation between TPR gap in an occupation and % of women in that occupation (lower is better). Each word dropout strategy outperforms BERT baseline on all three metrics and R-LACE/INLP on two metrics. D-BIAS achieves the best RMSE TPR gap.

UNIFORM: drops out words throughout the training vocabulary i.e., uniformly replaces those with `[MASK]` with a probability of $p = 0.3$.

SCRUB: replaces all mentions of identity words in the training vocabulary with `[MASK]`.

3.2 Results on Occupation Classification

We perform experiments on the scrubbed version of the data (with names and pronouns replaced with '_').[8] Following previous work, we use stratified-by-occupation splits, with 65% of the biographies for training, 10% for validation, and 25% for testing, resulting in $\sim$ 258K/40K/100K biographies for train, validation and test respectively (De-Arteaga et al., 2019). We evaluate our approaches in terms of TPR gap (difference between TPR for men and women biographies, lower is better) for each occupation (section 2). A high TPR gap indicates disparities in the model's ability to correctly classify biographies of men and women.

We present debiasing results in Table 1. In each case, we report means and variances (in parentheses) over 5 random initializations. The first row presents results using baseline BERT fine-tuned on the BIOS task without any dropout intervention. The next four rows show results using our proposed debiasing approaches. The last five rows present results on the baseline approaches (including R-LACE and INLP). The first column reports overall accuracy on the test set (higher is better). The second column presents the root mean square error (RMSE) of the TPR gap across the 28 occupations (lower is better). In the third column, following previous work, we compute the correlation ρ between the TPR gap in a given profession and the percentage of women in that profession (lower is better) following Ravfogel et al. (2022). This metric assesses the correlation between disparities in model predictions and existing disparities in the data.

In Table 1, each word dropout approach outperforms the fine-tuned BERT baseline on all three evaluation metrics. Moreover, all word dropout approaches also outperform R-LACE and INLP on both bias metrics (RMSE TPR gap and correlation). D-BIAS achieves the lowest (i.e. best) RMSE TPR gap, while only slightly decreasing classifier accuracy. SENT-ST achieves the second lowest RMSE followed by SENT-K. SENT-K outperforms R-LACE and INLP in terms of all three metrics. SENT-ST and SENT-K substantially improve the RMSE, while achieving test accuracies roughly equivalent to baseline BERT. D-BIAS and SENT-ST outperform SCRUB in lowering the RMSE. All word dropout approaches outperform UNIFORM in lowering RMSE and correlation (D-BIAST has slightly higher RMSE but the standard deviation is lower).

In Figure 1, we plot the TPR gap for each occupation, averaged over the 5 random initializations. D-BIAS clearly stands out, as it addresses the outliers (occupations such as `model` and `rapper` that have high TPR gaps). Moreover, it results in more

[8] https://github.com/microsoft/biosbias

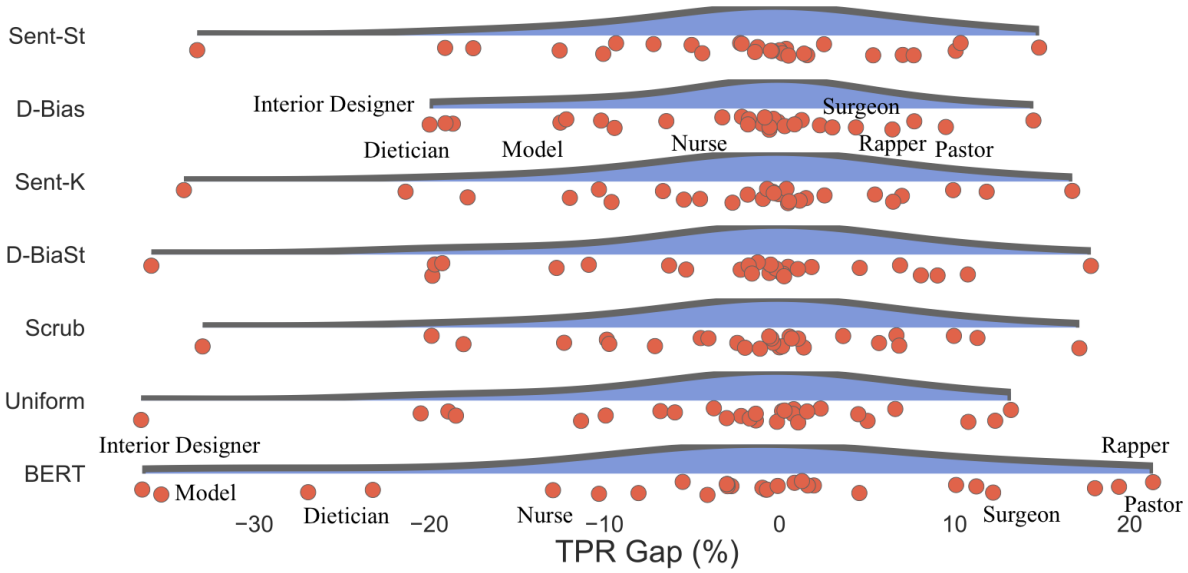

Figure 1: **Debiasing results on BIOS data as TPR gap** (averaged over 5 random initializations) for each occupation. For each method, red dots represent a strip plot of TPR gap across occupations; corresponding violin plot is in blue. D-BIAS outperforms other approaches; it decreases TPR gaps and achieves a more even distribution of TPR gaps across occupations.

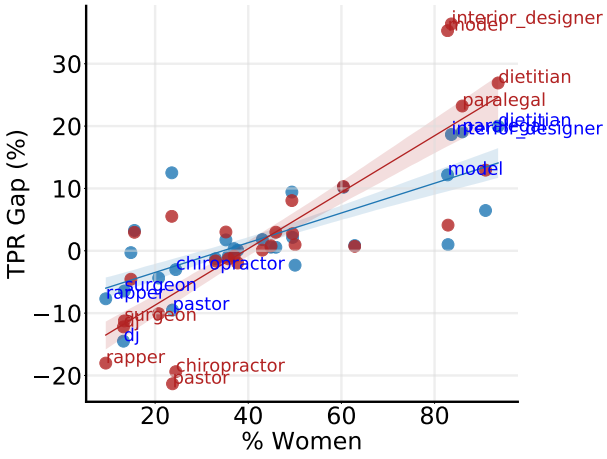

Figure 2: **TPR gap (averaged across 5 random initializations) vs. proportion of women across occupations**. The slope indicates correlation (lower is better) between representation disparities in training data and disparities in model predictions after fine-tuning. D-BIAS (blue) decreases correlation from 0.818 in BERT (red) to 0.728.

equitable TPR gap numbers across occupations, all of which are closer to 0 relative to TPR gaps with other approaches. We observe that D-BIAS substantially reduces the TPR gap across occupations (e.g. from $\sim 36\%$ to $\sim 12\%$ for `model`, from $\sim 21\%$ to $\sim 9.5\%$ for `pastor`).

In Figure 2, we plot the TPR gap in each occupation vs. the percentage of women in training data, in that occupation (ρ in Table 1 corresponds to the correlation between these two variables). We see that occupations such as `dietician`, `interior designer` and `model` tend to be dominated by women biographies, whereas `surgeon`, `rapper` and `chiropractor` are more likely to have biographies belonging to men. These occupations with skewed representation in the biographies show relatively large TPR gaps (corresponding to baseline BERT in red). The slope ($\rho \approx 0.8$ for BERT, Table 1) indicates that representation disparities in the fine-tuning data correlate with disparities in the model predictions. D-BIAS succeeds in reducing the TPR gap in these occupations. From Table 1, D-BIAS decreases the correlation to $\rho \approx 0.7$ whereas, $\rho \approx 0.8$ for both R-LACE and INLP.

3.3 Results on Toxicity Classification

Following the set up in Dixon et al. (2018), we divide the data into $\sim$ 96K comments for training (75%), $\sim$ 32K for validation and $\sim$ 32K for test. As in Dixon et al. (2018), we evaluate our approaches on a synthetic testset `madlibs` [9], with 89K examples created using templates of both toxic and non-toxic phrases that are filled in with a list of the 50 identity terms. Additionally, we evaluate on the held-out test set. Following previous work (Steed et al., 2022), we report downstream biases in terms

[9] https://github.com/conversationai/unintended-ml-bias-analysis/tree/main/unintended_ml_bias/eval_datasets

	Test Acc ↑	Group FPR (across identities)		Group Acc (across identities)	
		Mean ↓	Spread ↓	Mean ↑	Spread ↓
BERT	95.88 (4.15)	7.31 (1.13)	23.89 (1.66)	91.75 (0.64)	11.17 (0.78)
D-BIAS	96.59 (3.58)	**1.81 (1.63)**	**6.52 (5.21)**	93.31 (1.49)	**4.24 (1.86)**
D-BIAST	94.20 (5.90)	5.42 (1.57)	20.24 (2.92)	88.62 (1.46)	9.69 (1.05)
SENT-K	96.51 (3.53)	3.81 (0.57)	16.35 (0.78)	93.06 (0.67)	7.71 (0.26)
SENT-ST	**96.70 (3.31)**	5.03 (1.54)	18.92 (3.18)	**93.44 (0.33)**	9.03 (1.52)
UNIFORM	96.04 (4.06)	7.72 (0.71)	23.82 (1.19)	92.17 (1.19)	11.46 (0.50)
SCRUB	95.80 (4.27)	7.69 (1.01)	24.20 (1.94)	91.59 (1.01)	11.29 (0.87)

Table 2: **Debiasing results on WIKI `madlibs`, in terms of Group FPR and Group Accuracy**, averaged over 5 random initializations (variances in paranthesis). D-BIAS outperforms all approaches in terms of FPR (Group FPR Mean) and achieves an even distribution of FPR (Group FPR Spread) and accuracies (Group Acc Spread) across all identity groups.

	Test Acc ↑	Group FPR (across identities)		Group Acc (across identities)	
		Mean ↓	Spread ↓	Mean ↑	Spread ↓
BERT	98.43 (1.57)	3.55 (0.43)	4.33 (0.69)	90.68 (0.23)	6.41 (0.28)
D-BIAS	98.05 (1.95)	**1.69 (0.57)**	**3.31 (0.90)**	89.77 (0.55)	6.61 (0.58)
D-BIAST	98.32 (1.68)	3.52 (0.51)	4.87 (0.78)	90.15 (0.42)	6.59 (0.28)
SENT-K	98.41 (1.59)	3.58 (0.54)	4.43 (0.77)	90.56 (0.51)	**6.10 (0.29)**
SENT-ST	98.44 (1.57)	3.42 (0.36)	4.37 (0.70)	90.39 (0.42)	6.56 (0.22)
UNIFORM	98.39 (1.61)	4.87 (0.35)	6.07 (0.37)	**90.81 (0.50)**	6.16 (0.39)
SCRUB	**98.44 (1.56)**	3.44 (0.20)	4.17 (0.23)	90.66 (0.33)	6.38 (0.22)

Table 3: **Debiasing results on WIKI held-out testset, in terms of Group FPR and Group Accuracy**, for identity groups with ≥ 10 samples in testset, averaged over 5 random initializations (variances in parenthesis). D-BIAS outperforms all approaches in terms of FPR (Group FPR Mean and Spread), while achieving similar overall test accuracy and individual group accuracies.

of differences in FPR, across comments grouped according to mentions of identity words. A high FPR implies that the identity group is more likely to be flagged for associations with toxic content.

In Table 2, we present debiasing results on the `madlibs` testset, with means and variances (in parenthesis) over 5 random initializations. The first column presents overall test accuracy (higher is better). For each run, we compute a) mean group-wise accuracy/FPR (Group Acc Mean and Group FPR Mean respectively) across identities and b) standard deviation of the group-wise accuracy/FPR, which we denote as the *spread*. Ideally, we want higher Group Acc Mean, lower Group FPR Mean, and lower spread for both accuracy and FPR (since we do not want the model to exhibit large disparities in accuracy/FPR across the identity groups). Columns two to five present these numbers. The first row presents results with baseline BERT without any word dropout. The next rows present our proposed approaches, followed by the other baselines. We do not compare with R-LACE and INLP since those target binary bias attributes. We plot the distribution of group accuracy and group FPR (averaged over random seeds) in Figure 3.

We observe that BERT fine-tuned on the WIKI task exhibits significant variations in group accuracies and FPRs across the identity groups. The most glaring examples are the `gay` and `homosexual` identities. Anytime these identity words appear in a Talk Page comment, the model *always* predicts `Toxic`, irrespective of the context. This corresponds to 100% FPR for both these identity words (Figure 3) and the the relatively lower ($\sim 50\%$) prediction accuracies (Figure 6). We observe that all dropout approaches outperform BERT, UNIFORM and SCRUB.

D-BIAS substantially outperforms BERT without any dropout, for each evaluation metric. First, it improves model performance in terms of both

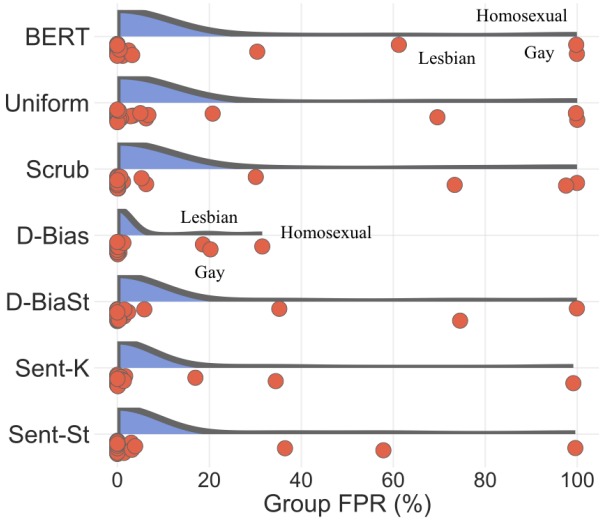

Figure 3: **Debiasing results on WIKI `madlibs` testset**: FPR for each identity group, averaged over 5 random initializations. Red dots show strip plots; corresponding violin plots are in blue. D-BIAS decreases FPR on `gay` and `homosexual` from $\sim 99.8\%$ (with BERT, for both) to $\sim 18\%$ and $\sim 30\%$ respectively.

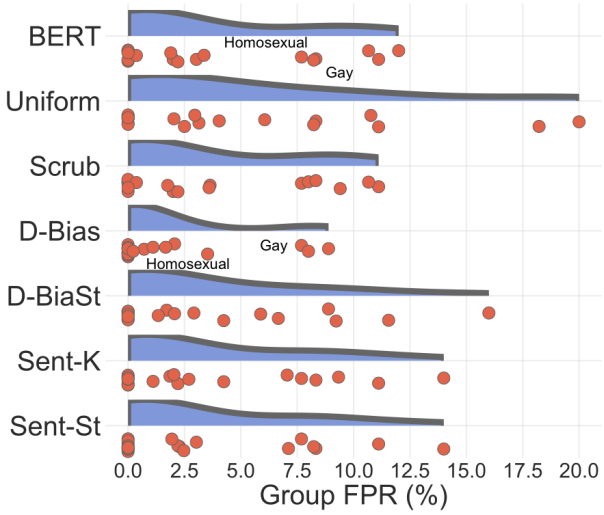

Figure 4: **Debiasing results on WIKI held-out testset**: FPR for identity group with at least 10 samples in test set, averaged over 5 random initializations. Red dots show strip plots; corresponding violin plots are in blue. On `homosexual`, D-BIAS decreases FPR from 8.2% in BERT to 3.5%.

overall test accuracy and the mean group accuracies. D-BIAS also decreases the mean group FPR (averaged across identities and random runs) from 7.31% to 1.81%. In particular for `gay`, average accuracy (over random initializations) remarkably increases to 83% (from $\sim 50\%$) and average FPR decreases to 38% (from 100%). Among the other word dropout approaches, SENT-K decreases Group FPR Mean, Group FPR Spread and Group Acc Spread the most. However, it does not address high FPR in case of the outlier identity categories such as `gay` and `homosexual`. SCRUB is largely ineffective, aligning with the findings in Steed et al. (2022). UNIFORM is also ineffective, highlighting the substantial advantage from using bias targeted word dropout strategies, compared to a uniform word dropout strategy.

Unlike `madlibs`, in which, all identities have uniform representation by construction, the held-out test set suffers from heavily skewed representation across identities (Figure 7). To eliminate noise, we report debiasing results for identity groups with at least 10 instances in this test set. In Table 3, we show that D-BIAS once again, outperforms all approaches in terms of FPR, while achieving similar overall test accuracy and group accuracies. In Figure 4, we highlight prominent reductions in downstream biases (FPR), e.g., for `homosexual`, from 8.2% in BERT to 3.5% with D-BIAS.

3.4 Analysis of nPMI Results

To better understand the nPMI scores in each case, we plot nPMI of commonly suspected proxy words (De-Arteaga et al., 2019) in Figure 5, with the `he/him` and `she/her` groups of pronouns. One notable observation about the gendered proxies is that the terms `husband` and `wife` have a stronger association with the opposite gender due to societal heteronormativity; these identities often appear in the form of 'her husband' and 'his wife'. Online biographies often include mentions of the author's family/personal life, in addition to occupation-related content. As a result, family-related words can easily act as proxies for gender information. Our nPMI framework scores these proxies reasonably well and, is able to find other gendered associations as well. For example, words such as `software`, `computers`, `technology` co-occur more with biographies of men (vs. `yoga` for women).

In case of the WIKI data, we present a list of the highest scoring nPMI words for a subset of the identity words, in Table 4. Notably, groups most likely to be flagged for toxic content, i.e., `gay`, `homosexual` frequently co-occur with abusive slang or are used in a pejorative sense, leading the models to associate toxicity with these identity groups.

Identity	Top-Ranking nPMI words
asian	afghans, persians, israelis, aryan, culturally, afghanistan, south-east, arabs
african	African-American, races, south, Civil, Obama, Africa, black, color, people
hispanic	phillipino, phillipinos, Spaniards, Spain, waves, Latin, Europe
indian	Bihar, valmiki, maharshi, subcontinent, goverment, Modi, hindi, Indus, singh
buddhist	buddhism, Asoka, patronizer, jainism, Guptas, mimansa, deities, edicts, monks
catholic	baptist, catholicism, nobility, christians, Pope, resignation, roman
muslim	tolerent, islam, divorce-divorce-divorce, jehad, balochistan-pakistan, ummah
jewish	humus, tautological, long-bearded, missionary, judaism, hebrew, jesus
gay	f******, d***, homophobia, same-gender, sucks, die, racist, lesbian, sexuality
homosexual	cross-gendered, masculinized, sexualorientation, transsexuals, abstention
queer	gender-binary, heteronormative, unconventional, insane, b****, a**, f***

Table 4: **List of top-ranking nPMI words for a subset of the identity groups in the WIKI data**. Groups such as gay and homosexual that are most commonly flagged for toxic content in model predictions, frequently co-occur with abusive slang or are used in a pejorative sense, leading models to associate toxicity with these identity words.

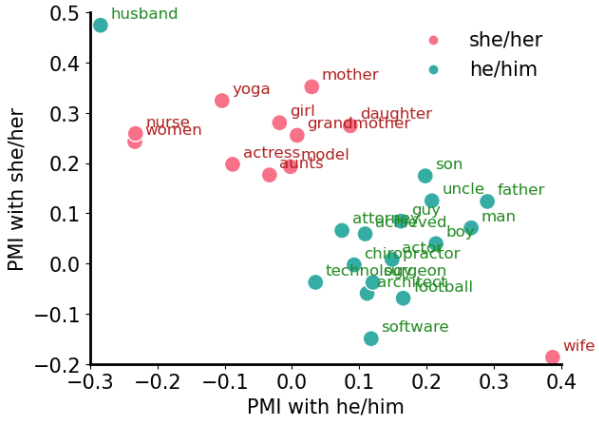

Figure 5: **nPMIs computed on BIOS for common proxy words**, with he/him and she/her groups of pronouns. Our nPMI framework scores common proxies reasonably well and also finds other gendered associations in the data, shown in red (for she/her) and green (for he/him) clusters of highly co-occurring words in each case.

4 Related Work

Prior work on de-biasing pre-trained language models largely focus on upstream mitigation of intrinsic biases (Meade et al., 2022; Kaneko and Bollegala, 2021; Schick et al., 2021). However, recent findings (Goldfarb-Tarrant et al., 2021; Steed et al., 2022) suggest that debiasing effects from upstream mitigation *may not* hold downstream, motivating the need for interventions focused on the context of downstream tasks.

(Ravfogel et al., 2022, 2020; Liang et al., 2020) explore de-biasing contextual representations by identifying and subtracting away, a linear subspace from the embeddings space. These approaches aim to identify one or several sets of linear subspaces that most accurately describe biased artifacts, such as gender biases. These approaches work with task-specific data, by adapting a linear layer on top of the debiased representations (post fine-tuning, after a bias subspace is subtracted from the representations). However, these approaches largely focus on binary attributes for bias and can often be computationally intensive. These can either require massive text corpora to construct templates for bias subspace generation (Liang et al., 2020) or use contrastive learning for debiasing (Cheng et al., 2021), also relying on massive external text corpora for creating augmented examples, or require iterative optimization to identify large sets of linear subspaces (Ravfogel et al., 2022, 2020)). In contrast, our approach is relatively straightforward to implement, does not make linearity assumptions, and easily extends to multiple identities.

5 Conclusions

In this paper, we show, in the context of two classification tasks, that eliminating proxies for identity words, in the fine-tuning data can substantially reduce downstream biases. Our findings underscore the importance of targeted and context-specific debiasing approaches, with a focus on attenuating stereotypical associations in the fine-tuning data.

Limitations

In this section, we outline some of the limitations of our approach. First, we focus our experiments only

on two downstream tasks: a) toxicity classification and b) occupation classification. As a result, our findings may not hold for all tasks, especially in non-classification tasks, where the information loss from removing proxies may more strongly impact performance.

Our results may also not hold for all kinds of stereotypical associations or demographic/social identity representations, especially as our method relies on PMIs to identify proxies based on co-occurrence statistics. There may be obvious proxies, for example, that are not based on co-occurence, but rather substitution for a particular identity term. PMI calculations also tend to be dominated by sparse co-occurences, so it is important to appropriately threshold how proxies are selected. Finally, we identify proxies at the token-level, but there may be proxies that only become apparent at larger n-grams.

Furthermore, it is not straightforward to account for various types of biases arising in different contexts. For example, the word gay could be used to indicate identity information or it can be used in a pejorative sense. However, our findings are encouraging and can inspire practitioners to focus more on the particular statistics and context of their fine-tuning data.

We perform all of our fine-tuning experiments with BERT, in order to achieve consistent evaluations with previous work in this space. While our approach can work with other pre-trained models, the results may not generalize. However, we are optimistic that debiasing effects will continue to hold, as many of the most commonly used pre-trained models use similar architectures and pre-training strategies as BERT.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter universal dependency parsing for african-american and mainstream american english. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

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

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Association for Computational Linguistics*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

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

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *arXiv preprint 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.

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.

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.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language understanding paper.pdf*.

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.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

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.

Appendix

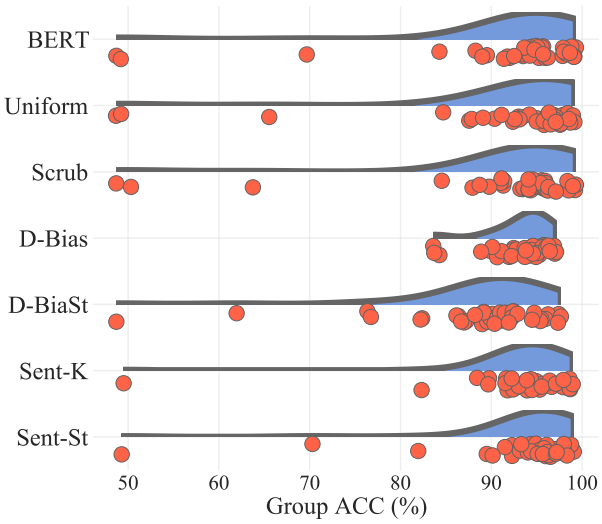

Figure 6: **Debiasing results on WIKI `madlibs` testset**: Test accuracy for each identity group, averaged over 5 random initializations. Red dots show the strip plots; corresponding violin plots are in blue. D-BIAS outperforms other approaches in achieving higher group accuracies, and a more equitable distribution of the accuracies across identity groups.

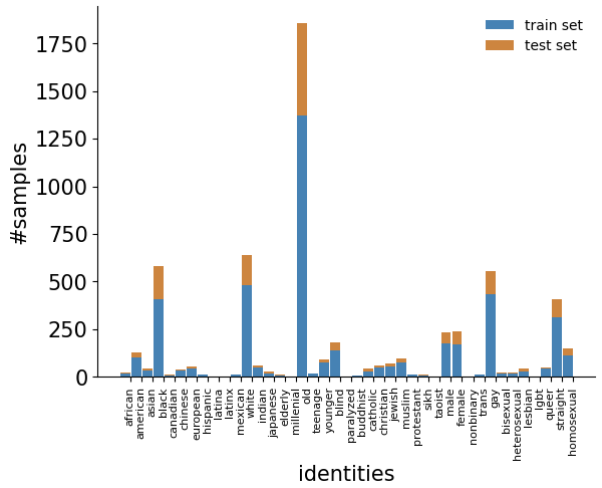

Figure 7: **Number of Talk Page Comments with mentions of each identity**, in the train and test splits of the WIKI data, demonstrating skewed representation across identoty groups.

6 Reproducibility Criteria

We fine-tune BERT-base-uncased for 5 fine-tuning epochs in each experiment, saving checkpoints after every epoch. We pass the `[CLS]` token representation through a binary or multi-class classification head for toxicity classification and occupation classification respectively. We choose the best checkpoint based on validation set accuracy. We set batch size to 32 for both training and evaluation. We trim the input text in each case to a maximum length of 128 tokens. We use Adam optimizer with a learning rate of 1e-05. Each fine-tuning experiment for 5 epochs (including training, validation and test) takes a little more than 1 hour for toxicity classification and about 3 hours for occupation classification on an NVIDIA Tesla A100 16 GB GPU.

- WIKI data is available at https://github.com/conversationai/unintended-ml-bias-analysis.
- The `madlibs` test set is available at https://github.com/conversationai/unintended-ml-bias-analysis/blob/main/archive/unintended_ml_bias/eval_datasets/bias_madlibs_89k.csv.
- BIOS data is available at: https://github.com/microsoft/biosbias.

In Table 1, we report results with R-LACE and INLP from Table 2 in Ravfogel et al. (2022).