How Gender Debiasing Affects Internal Model Representations, and Why It Matters

Hadas Orgad[1] **Seraphina Goldfarb-Tarrant**[2] **Yonatan Belinkov**[1]*

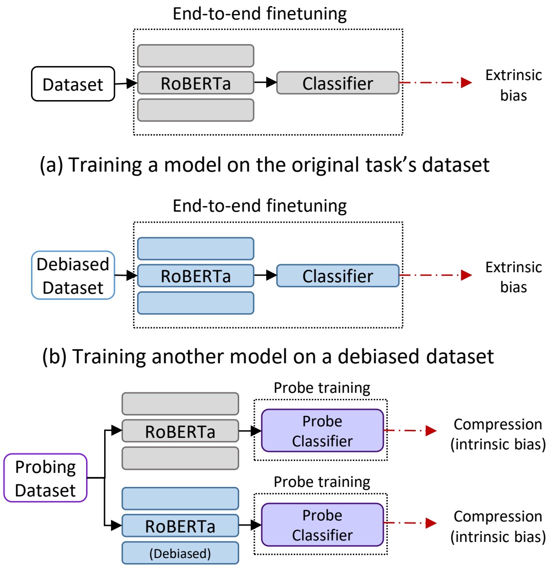
[1]Technion – Israel Institute of Technology [2]University of Edinburgh
orgad.hadas@cs.technion.ac.il s.tarrant@ed.ac.uk
belinkov@technion.ac.il

Abstract

Common studies of gender bias in NLP focus either on extrinsic bias measured by model performance on a downstream task or on intrinsic bias found in models' internal representations. However, the relationship between extrinsic and intrinsic bias is relatively unknown. In this work, we illuminate this relationship by measuring both quantities together: we debias a model during downstream fine-tuning, which reduces extrinsic bias, and measure the effect on intrinsic bias, which is operationalized as bias extractability with information-theoretic probing. Through experiments on two tasks and multiple bias metrics, we show that our intrinsic bias metric is a better indicator of debiasing than (a contextual adaptation of) the standard WEAT metric, and can also expose cases of superficial debiasing. Our framework provides a comprehensive perspective on bias in NLP models, which can be applied to deploy NLP systems in a more informed manner. [1]

1 Introduction

Efforts to identify and mitigate gender bias in Natural Language Processing (NLP) systems typically target one of two notions of bias. *Extrinsic* evaluation methods and debiasing techniques focus on the bias reflected in a downstream task (De-Arteaga et al., 2019; Zhao et al., 2018), while *intrinsic* methods focus on a model's internal representations, such as word or sentence embedding geometry (Caliskan et al., 2017; Bolukbasi et al., 2016; Guo and Caliskan, 2021). Despite an abundance of evidence pointing towards gender bias in pretrained language models (LMs), the extent of harm caused by these biases is not clear when it is not reflected in a specific downstream task (Barocas et al., 2017; Kate Crawford, 2017; Blodgett et al., 2020; Bommasani et al., 2021). For instance, while the word embedding proximity of "doctor" to "man" and "nurse" to "woman" is intuitively normatively wrong, it is not clear when such phenomena would lead to downstream predictions manifesting in social biases. Recently, Goldfarb-Tarrant et al. (2021) have shown that debiasing static embeddings intrinsically is not correlated with extrinsic gender bias measures, but the nature of the reverse relationship is unknown: how are extrinsic interventions reflected in intrinsic representations? Furthermore, Gonen and Goldberg (2019a) demonstrated that a number of intrinsic debiasing methods applied to static embeddings only partially remove the bias and that most of it is still hidden within the embed-



(c) Measuring our intrinsic metric on the debiased and original dataset using probing

Figure 1: Our proposed framework. Black arrows mark forward passes, red arrows mark things we measure. We first (a) train a model on a downstream task, then (b) train another model on the same task using a debiased dataset, and finally (c) measure intrinsic bias in both models and compare.

[*] Supported by the Viterbi Fellowship in the Center for Computer Engineering at the Technion.

[1] Our code and model checkpoints are publicly available at https://github.com/technion-cs-nlp/gender_internal

ding. Complementing their view, we examine *extrinsic* debiasing methods, as well as demonstrate the possible harm this could cause. Contrary to their conclusion, we do not claim that these debiasing methods should not be trusted, *as long as they are utilized with care*.

Our goal is to gain a better understanding of the relationship between a model's internal representations and its extrinsic gender bias by examining the effects of various debiasing methods on the model's representations. Specifically, we fine-tune models with and without gender debiasing strategies, evaluate their external bias using various bias metrics, and measure intrinsic bias in the representations. We operationalize intrinsic bias via two metrics: First, we use CEAT (Guo and Caliskan, 2021), a contextual adaptation of the widely used intrinsic bias metric WEAT (Caliskan et al., 2017). Second, we propose to use an information-theoretic probe to quantify the degree to which gender can be extracted from the internal model representations. Then, we examine how these intrinsic metrics correlate with a variety of extrinsic bias metrics that we measure on the model's downstream performance. Our approach is visualised in Figure 1.

We perform extensive experiments on two downstream tasks (occupation prediction and coreference resolution); several debiasing strategies that involve alterations to the training dataset (such as removing names and gender indicators, or balancing the data by oversampling or downsampling); and a multitude of extrinsic bias metrics. Our analysis reveals new insights into the way language models encode and use information on gender:

- The effect of debiasing on internal representations is reflected in gender extractability, while not always in CEAT. Thus, gender extractability is a more reliable indicator of gender bias in NLP models.
- In cases of high gender extractability but low extrinsic bias metrics, the debiasing is superficial, and the internal representations are a good indicator for this: The bias is still present in internal representations and can be restored by retraining the classification layer. Therefore, our proposed measuring method can help in detecting such cases before deploying the model.
- The two tasks show different patterns of correlation between intrinsic and extrinsic bias. The coreference task exhibits a high correlation. The occupation prediction task exhibits a lower correlation, but it increases after retraining (a case of superficial debiasing). Gender extractability shows higher correlations with extrinsic metrics than CEAT, increasing the confidence in this metric as a reliable measure for gender bias in NLP models.

2 Methodology

In this study, we investigate the relationship between extrinsic bias metrics of a task and a model's internal representations, under various debiasing conditions, for two datasets in English. We perform extrinsic debiasing, evaluate various extrinsic and intrinsic bias metrics before and after debiasing, and examine correlations.

Dataset. Let $D = \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ be a dataset consisting of input data $\boldsymbol{X}$, labels $\boldsymbol{Y}$ and protected attributes $\boldsymbol{Z}$.[2] This work focuses on gender as the protected attribute z. In all definitions, F and M indicate female and male gender, respectively, as the value of the protected attribute z.

Trained Model. The model is optimized to solve the downstream task posed by the dataset. It can be formalized as $f \circ g : \boldsymbol{X} \rightarrow \mathbb{R}^{|\mathcal{Y}|}$, where $g(\cdot)$ is the feature extractor, implemented by a language model, e.g., RoBERTa (Liu et al., 2019), $f(\cdot)$ is the classification function, and $\mathcal{Y}$ is the set of the possible labels for the task.

2.1 Bias Metrics

Each bias evaluation method described in the literature can be categorized as extrinsic or intrinsic. In all definitions, $\mathbf{r}$ indicates the model's output probabilities.

2.1.1 Extrinsic Metrics

Extrinsic methods involve measuring the bias of a model solving a downstream problem. The extrinsic metric is a function:

$$E(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R}, \boldsymbol{Z}) \in \mathbb{R}$$

The output represents the quantity of bias measured; the further from 0 the number is, the larger the bias is. Our analysis comprises a wide range

[2] $\boldsymbol{Z}$ is by convention used for attributes for which we want to ensure fairness, such as gender, race, etc. It is purposefully broad, and depending on the task and data could refer to the gender of an entity in coreference, the subject of a text, the demographics of the author of a text, etc.

of extrinsic metrics, including some that have been measured in the past on the analyzed tasks (Zhao et al., 2018; De-Arteaga et al., 2019; Ravfogel et al., 2020; Goldfarb-Tarrant et al., 2021) and some that have never been measured before, and shows our results apply to many of them. For illustration, we will consider occupation prediction, a common task in research on gender bias (De-Arteaga et al., 2019; Ravfogel et al., 2020; Romanov et al., 2019). The input x is a biography and the prediction y is the profession of the person described in it. The protected attribute z is the gender of that person.

Performance gap. This is the difference in performance metric for two different groups, for instance two groups of binary genders, or a group of pro-stereotypical and a group of anti-stereotypical examples. We measure the following metrics: True Positive Rate (TPR), False Positive Rate (FPR), and Precision. In occupation prediction, for instance, the TPR gap for each profession y expresses the difference in the percentage of women and men whose profession is y and are correctly classified as such. We also measure F1 of three standard clustering metrics for coreference resolution. Each such performance gap captures a different facet of gender bias, and one might be more interested in one of the metrics depending on the application.

We compute two types of performance gap metrics: (1) the sum of absolute gap values over all classes; (2) the Pearson correlation between the performance gap for a class and the percentage of women in that class. For instance, if y is a profession, we measure the correlation between performance gaps and percentages of women in each profession.[3] The two metrics are closely related but answer slightly different questions: the sum quantifies how a model behaves differently on different genders, and the correlation shows the relation of model behaviour to social biases (in the world or the data) without regard to actual gap size.

Statistical metrics. For breadth of analysis, we examine three additional statistical metrics (Barocas et al., 2019), which correspond to different notions of bias. All three are measured as differences (d) between two probability distributions, and we then obtain a single bias quantity per metric by summing all computed distances.

- *Independence*: $d\big(P(\mathbf{r}|\mathbf{z}=z), P(\mathbf{r})\big)\forall z \in \{F, M\}$. For instance, we measure the difference between the distribution of model's predictions on women and the distribution of all predictions. Independence is stronger as the prediction $\mathbf{r}$ is less correlated with the protected attribute $\mathbf{z}$. It is measured with no relation to the gold labels.
- *Separation*: $d\big(P(\mathbf{r}|\mathbf{y}=y,\mathbf{z}=z), P(\mathbf{r}|\mathbf{y}=y)\big)$ $\forall y \in \mathcal{Y}, z \in \{F, M\}$. For instance, we measure the difference between the distribution of a model's predictions on women who are teachers and the distribution of predictions on all teachers. It encapsulates the TPR and FPR gaps discussed previously, and can be seen as a more general metric.
- *Sufficiency*: $d\big(P(\mathbf{y}|\mathbf{r}=r,\mathbf{z}=z), P(\mathbf{y}|\mathbf{r}=r)\big)$. For instance, we measure the difference between the distribution of gold labels on women classified as teachers by the model and the distribution of gold labels on all individuals classified as teachers by the model. Sufficiency relates to the concept of calibration in classification. A difference in the classifier's scores for men and for women indicates that it might be penalizing or over-promoting one of the genders.

2.1.2 Intrinsic Metrics

Intrinsic methods are applied to the representation obtained from the feature extractor. These methods are independent of any downstream task. The intrinsic metric is a function:

$$I(g(\boldsymbol{X}), \boldsymbol{Z}) \in \mathbb{R}$$

Compression. Our main intrinsic metric is the *compression* of gender information evaluated by a minimum description length (MDL) probing classifier (Voita and Titov, 2020), trained to predict gender from the model's representations. Probing classifiers are widely used for predicting various properties of interest from frozen model representations (Belinkov and Glass, 2019). MDL probes were proposed because a probe's accuracy may be misleading due to memorization and other issues (Hewitt and Liang, 2019; Belinkov, 2021). We use the MDL online code, where the probe is trained in timesteps, on increasing subsets of the training set, then evaluated against the rest of it. Higher compression indicates greater gender extractability.

CEAT. We also measure CEAT (Guo and Caliskan, 2021), which is a contextualized version

[3] Percentages for coreference resolution are taken from labour statistics, following Zhao et al. (2018). For occupation prediction we use training set statistics following De-Arteaga et al. (2019), *before* balancing.

of WEAT (Caliskan et al., 2017), a widely used bias metric for static word embeddings. WEAT defines sets $\mathbb{X}$ and $\mathbb{Y}$ of target words, and sets $\mathbb{A}$ and $\mathbb{B}$ of attribute words. For instance, $\mathbb{A}$ and $\mathbb{B}$ contain males and females names, while $\mathbb{X}$ and $\mathbb{Y}$ contain career and family related words, respectively. The bias is operationalized as the geometric proximity between the target and attribute word embeddings, and is quantified in CEAT by the Combined Effect Size (CES) and a p-value for the null hypothesis of having no biased associations. For more information on CEAT refer to Appendix A.4.3.

2.2 Debiasing Techniques

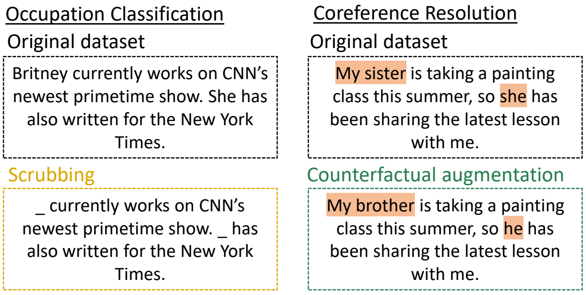
We debias models by modifying the downstream task's training data before fine-tuning. *Scrubbing* (De-Arteaga et al., 2019) removes first names and gender-specific terms ("he", "she", "husband", "wife", "Mr", "Mrs", etc.). *Balancing* subsamples or oversamples examples such that each gender is equally represented in the resulting dataset w.r.t each label. *Anonymization* (Zhao et al., 2018) removes named entities. *Counterfactual Augmentation* (Zhao et al., 2018) involves replacing male entities in an example with female entities, and adding the modified example to the training set. As some of these are dataset/task-specific, we give more details in the following section.

3 Experiments

In each experiment, we fine-tune a model for a downstream task. For training, we use either the original dataset or a dataset debiased with one of the methods from Section 2.2. Figure 2 presents examples of debiasing methods for the two downstream tasks. We measure two intrinsic metrics by probing that model's inner representations for gender extractability (as measured by MDL) and by CEAT, and test various extrinsic metrics. The relation between one intrinsic and one extrinsic metric becomes one data point, and we repeat over many random seeds (for both the model and the probe). Further implementation details are in appendix A.

3.1 Occupation Prediction

The task of occupation prediction is to predict a person's occupations (from a closed set), based on their biography. We use the Bias in Bios dataset (De-Arteaga et al., 2019). Regardless of the training method, the test set is subsampled such that each profession has equal gender representation.



Britney currently works on CNN's newest primetime show. She has also written for the New York Times.

_ currently works on CNN's newest primetime show. _ has also written for the New York Times.

My sister is taking a painting class this summer, so she has been sharing the latest lesson with me.

My brother is taking a painting class this summer, so he has been sharing the latest lesson with me.

Figure 2: Examples of two debiasing methods performed on the data.

Model. Our main model is a RoBERTa model (Liu et al., 2019) topped with a linear classifier, which receives the [CLS] token embedding as input and generates a probability distribution over the professions. In addition, we experiment with training a baseline classifier layer on top of a frozen, non-finetuned RoBERTa. We also replicate our RoBERTa experiments with a DeBERTa model (He et al., 2020), to verify that our results are are not model specific and hold more broadly.

Debiasing Techniques. Following De-Arteaga et al. (2019) we experiment with scrubbing the training dataset. Figure 2 shows an example biography snippet and its scrubbed version. We also conduct balancing (per profession, subsampling and oversampling to ensure an equal number of males and females per profession), which has not previously been used on this dataset and task.

Metrics. We measure all bias metrics from Section 2.1 except for F1.

Probing. The probing dataset for this task is the test set, and the gender label of a single biography is the gender of the person described in it. We probe the [CLS] token representation of the biography. In addition to the models described above, we measure baseline extractability of gender information from a randomly initialized RoBERTa model.

3.2 Coreference Resolution

The task of coreference resolution is to find all textual expressions referring to the same real-world entities. We train on Ontonotes 5.0 (Weischedel et al., 2013) and test on the Winobias challenge dataset (Zhao et al., 2018). Winobias consists of sentence pairs, pro- and anti-stereotypical variants, with individuals referred to by their profession. For example, "The physician hired the secretary be-

Debiasing Strategy	Intrinsic		Extrinsic							
			Before				After			
	Compression	CEAT	TPR (P)	FPR (S)	Sep	Suff	TPR (P)	FPR (S)	Sep	Suff
Random	5.61*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	10.12	0.49*	-	-	-	-	-	-	-	-
None	4.12	0.22	0.76	0.08	0.33	9.45	0.78	0.073	0.33	9.70
Oversampling	8.52*	0.29	0.73	0.09*	0.31	8.32*	0.81*	0.068*	0.33	10.91*
Subsampling	3.57	0.22	**0.32***	**0.03***	**0.20***	**1.22***	**0.70***	0.08*	0.30*	1.32*
Scrubbing	**1.70***	0.23	0.70*	0.06*	0.30	4.93*	0.71*	**0.06***	**2.56***	**0.81***

(a) Occupation classification: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and per retrained classification model.

Debiasing Strategy	Intrinsic		Extrinsic							
			Before				After			
	Compression	CEAT	F1 diff	FPR (S)	Sep	Suff	F1 diff	FPR (S)	Sep	Suff
Random	0.83*	0.12†	-	-	-	-	-	-	-	-
Pre-trained	0.96	0.49*	-	-	-	-	-	-	-	-
None	1.98	0.35	6.63	0.12	1.25	8.69	6.07	0.11	1.19	7.35
Anon	2.07*	0.31*	7.26	0.13	1.34	8.82	7.42*	0.13*	1.34*	8.66*
CA	**1.50***	0.27*	**2.30***	0.05*	**0.54***	1.67*	3.67*	0.06*	0.67*	2.40*
Anon + CA	1.54*	**0.25***	2.42*	**0.049***	0.56*	**1.56***	**2.86***	**0.05***	**0.59***	**1.65***

(b) Coreference resolution: Comparison of intrinsic and extrinsic metrics before and after retraining of classification layer, over 10 seeds per fine-tuned model and 5 seeds per retrained classification model.

Table 1: Results on both tasks. * marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score in each column is marked with **bold**. P = Pearson; S = Sum. † was computed only on 3 out of 10 tests for which CEAT's $p < 0.05$.

cause *he/she* was busy." is pro/anti-stereotypical, based on US labor statistics. [4] A coreference system is measured by the performance gap between the pro- and anti-stereotypical subsets.

Model. We use the model presented in Lee et al. (2018a) with RoBERTa as a feature extractor.

Debiasing Techniques. Following Zhao et al. (2018), we apply anonymization (denoted as Anon) and counterfactual augmentation (CA) on the training set. These techniques were used jointly in previous work; we examine each individually as well.

Metrics. Following Zhao et al. (2018), we measure the F1 difference between anti- and pro-stereotypical examples.[5] We also interpret the task as a classification problem, and measure all metrics from Section 2.1. For more details refer to Appendix A.4.2.

Probing. We probe the representation of a profession word as extracted from Winobias sentences, after masking out the pronouns. We define a profession's gender as the stereotypical gender for this profession. To prevent memorization by the probe—given the small number of professions—the dataset is sorted so that professions are gradually added to the training set, so a success on the validation set is on previously unseen professions.

4 Results

Tables 1a and 1b present intrinsic and extrinsic metrics for RoBERTa models on the occupation prediction and coreference resolution tasks, respectively. We present a representative subset of the measured metrics that demonstrate the observed phenomena; full results are found in Appendix B. The DeBERTa model results are consistent with the RoBERTa model trends.

4.1 Compression Reflects Debiasing Effects

As shown in the tables, compression captures differences in models that were debiased differently. CEAT, however, cannot differentiate between occupation prediction models. For example, in occupation prediction (Table 1a) the compression rate

[4] Labor Force Statistics from the Current Population Survey, https://www.bls.gov/cps/cpsaat11.htm

[5] We combined the T1 and T2 datasets, as well as the dev and test datasets, to create a single held-out challenge dataset.

varies significantly between a non-debiased and a debiased model via scrubbing and oversampling, while CEAT detects no difference between the models. In coreference resolution (Table 1b), both compression and CEAT are able to identify differences between the non-debiased model and the others, such as CA, which has both a lower compression and CEAT effect. But the CEAT effect sizes are small (below 0.5), which implies no bias, in contrast to the extrinsic metrics.

4.2 High Gender Extractability Implies Superficial Debiasing

Extrinsic and intrinsic effects of debiasing. In occupation classification (Table 1a), somewhat surprisingly, subsampling the training data has the strongest effect on extrinsic metrics, but not on compression rate. Scrubbing reduces both intrinsic and extrinsic metrics, although its effect on extrinsic metrics is limited compared to subsampling. Training with oversampling caused less reduction in extrinsic bias metrics. A consequence of oversampling is that some metrics are less biased, but compression rates are increased, so gender information is more accessible. The effectiveness of subsampling over other metrics is further discussed in appendix C. In coreference resolution (Table 1b), while both CA and CA with anonymization reduced gender extractability as well as external bias metrics, anonymization alone *increased* intrinsic bias without affecting external bias metrics significantly.

Debiasing without fine-tuning. As the effect on extrinsic bias did not match the effect on intrinsic bias in several cases, we examined the role of the classification layer. We trained a model for occupation prediction without fine-tuning the underlying RoBERTa model. Training on a subsampled dataset also reduced the extrinsic metrics (0.15, 0.03, 0.20, and 0.31, respectively, on TPR gaps Pearson, FPR gaps sum, separation sum, and sufficiency sum). Detailed results of this experiment can be found in Appendix B. Since no updates were made to the LM, the internal representations could not be debiased, thus the debiasing observed in this model can only be superficial.

Retraining the classification layer. Fine-tuning of both tasks revealed that lower extrinsic metrics did not always lead to lower compression. Does this indicate cases where the debiasing process is only superficial, and the internal representations remain biased? To test this hypothesis, we froze the previously fine-tuned LM's weights, and retrained the classification layer. We used the original (non-debiased) training set for retraining.

Tables 1a and 1b also compare extrinsic metrics before and after retraining. All models show bias restoration, due to the classification layer being trained on the biased dataset.[6] The amount of bias restored varies between models in a way that is predictable by the compression metric.

In the occupation prediction task, comparing Before and After numbers in Table 1a, the model fine-tuned using a scrubbed dataset—which has the lowest compression rate—displays the least bias restoration, confirming that the LM absorbed the process of debiasing. The model fine-tuned on subsampled data has higher extrinsic bias after retraining. Hence, the debiasing was primarily cosmetic, and the representations within the LM were not debiased. The model fine-tuned on oversampled data—which has the highest compression—has the highest extrinsic bias (except for FPR), even though this was not true before retraining.

In coreference resolution, comparing Before and After numbers in Table 1b, models with the least extrinsic bias (CA and CA+Anon) are also least biased after retraining. Compression rate predicted this; these models also had lower compression rates than non-debiased models. Interestingly, the model fine-tuned with an anonymized dataset is the most biased after retraining, consistent with its high compression rate relative to the other models. As with subsampling and oversampling in occupation prediction, anonymization's (lack of) effect on extrinsic metrics was cosmetic (compare None and Anon in Before block, Table 1b). Anonymization actually had a biasing effect on the LM, which was realized after retraining.

We conclude that compression rate is a useful indicator of superficial debiasing, and can potentially be used to verify and gain confidence in attempts to debias an NLP model, especially when there is little or no testing data.

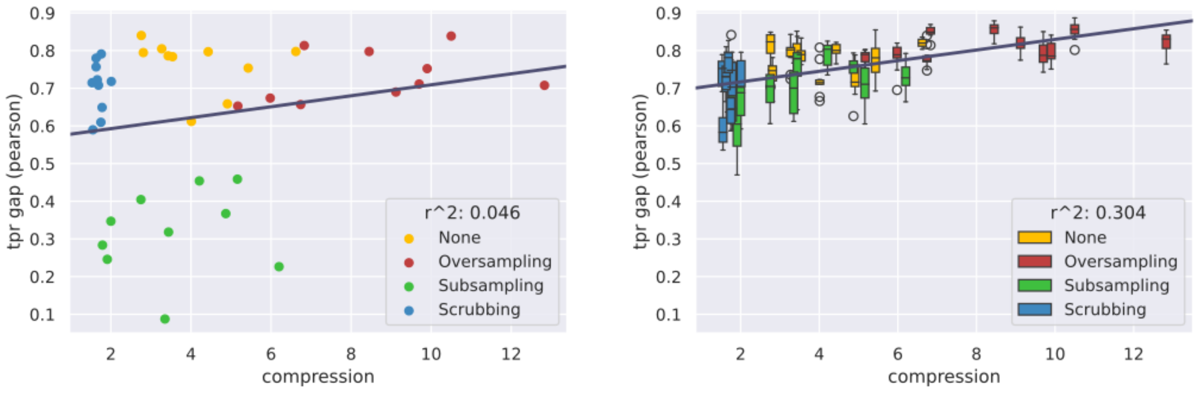
4.3 Correlation between Extrinsic and Intrinsic Metrics

Table 2 shows correlations between compression rate and various extrinsic metrics before and after

[6] The training datasets contain bias. The occupation prediction set has an unbalanced amount of males and females per profession (for example 15% of software engineers are females). The coreference resolution training set has more male than female pronouns, and males are more likely to be referred to by their profession (Zhao et al., 2018).

	Occupation Classification				**Coreference Resolution**			
	R^2 Compression		R^2 CEAT		R^2 Compression		R^2 CEAT	
Metric	Before	After	Before	After	Before	After	Before	After
F1 diff ($pro - anti$)	-	-	-	-	0.821	0.709	0.246	0.005
TPR gap (P)	0.046	0.304	0.042	0.049	0.222	0.006	0.008	0.012
TPR gap (S)	0.049	0.449	0.022	0.036	0.817	0.752	0.297	0.003
FPR gap (P)	0.001	0.120	0.008	0.002	0.021	0.054	0.002	0.000
FPR gap (S)	0.353	0.046	0.079	0.001	0.844	0.773	0.263	0.004
Precision gap (P)	0.004	0.063	0.006	0.002	0.223	0.008	0.009	0.013
Precision gap (S)	0.150	0.291	0.031	0.054	0.817	0.752	0.296	0.003
Independence gap (S)	0.251	0.382	0.050	0.005	0.778	0.732	0.355	0.001
Separation gap (S)	0.066	0.165	0.046	0.009	0.835	0.776	0.261	0.005
Sufficiency gap (S)	0.202	0.567	0.040	0.034	0.825	0.753	0.287	0.002

Table 2: Coefficient determination of the regression line taken on the compression rate or CEAT and each extrinsic metric, before and after retraining of the classification layer. P = Pearson; S = Sum.



(a) Fine-tuned models. Each point is a single seed for training and testing the model.

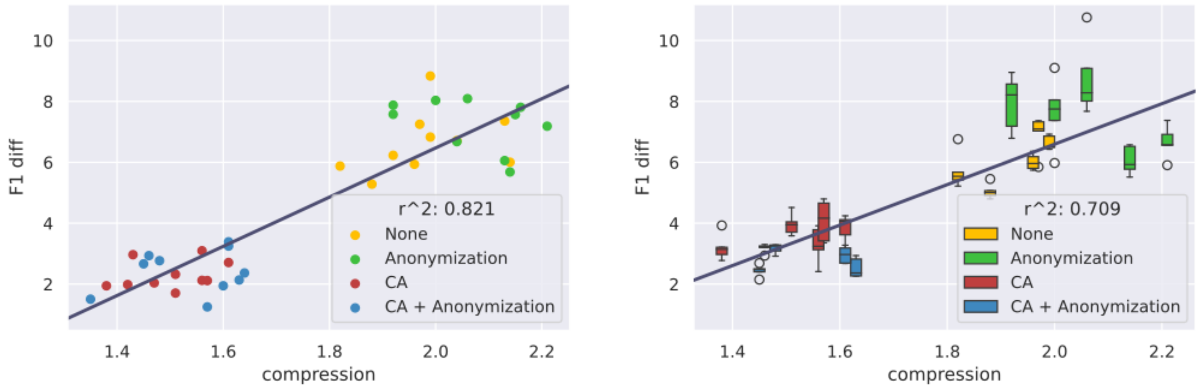
(b) After retraining. Each box represents 10 runs of retraining on the same fine-tuned feature extractor.

Figure 3: Occupation prediction: Compression vs. TPR-gap (Pearson) after various debiasing strategies.

retraining. In occupation prediction, certain extrinsic metrics have a weak correlation with compression rate, while others do not. Except one metric (FPR gap sum), the compression rate and the extrinsic metric correlate more after retraining. Figure 3 illustrates this for TPR-gap (Pearson). The increase is due to superficial debiasing, especially by subsampling data, which prior to retraining had low extrinsic metrics and relatively high intrinsic metrics. This shows that correlation between extrinsic metrics and compression rate for certain metrics is stronger than it appeared before retraining. It is unsurprising that CEAT does not correlate with any extrinsic metrics, since CEAT could not distinguish between different models.

Coreference resolution shows stronger correlations between compression rate and extrinsic metrics, but low correlations between Pearson metrics. We further discuss cases of no correlation in appendix D. Correlations decrease after retraining, but metrics that were highly correlated remain so (> 0.7 after retraining). The correlations are visualized for F1 difference metrics in Figure 4. CEAT and extrinsic metrics correlate much less than compression rate (Table 2). Our results are in line with those of Goldfarb-Tarrant et al. (2021), who found a lack of correlation between extrinsic metrics and WEAT, the static-embedded version of CEAT.

Given that recent work (Goldfarb-Tarrant et al., 2021; Cao et al., 2022) questions the validity of intrinsic metrics as a reliable indicator for gender bias, the compression rate provides a reliable alternative to current intrinsic metrics, by offering correlation to many extrinsic bias metrics.



(a) Fine-tuned models. Each point is a single seed for training and testing the model.

(b) After retraining. Each box represents 5 runs of retraining on the same fine-tuned feature extractor.

Figure 4: Coreference resolution: Compression vs. F1 difference after various debiasing strategies.

5 Related Work

There are few studies that examine both intrinsic and extrinsic metrics. Previous work by Goldfarb-Tarrant et al. (2021) showed that debiasing static embeddings intrinsically is not correlated with extrinsic bias, challenging the assumption that intrinsic metrics are predictive of bias. We examine the other direction, exploring how extrinsic debiasing affects intrinsic metrics. We also extend beyond their work to contextualized embeddings, a wider range of extrinsic metrics, and a new, more effective intrinsic metric based on information-theoretic probing. A contemporary work by Cao et al. (2022) measured the correlations between intrinsic and extrinsic metrics in contextualized settings across different language models. In contrast, our work examines the correlations across different versions of the same language model by fine-tuning it using various debiasing techniques.

Studies that inspect extrinsic metrics include either a challenge dataset curated to expose differences in model behavior by gender, or a test dataset labelled by gender. Among these datasets are Winobias (Zhao et al., 2018), Winogender (Rudinger et al., 2018) and GAP (Webster et al., 2018) for coreference resolution, WinoMT (Stanovsky et al., 2019) for machine translation, EEC (Kiritchenko and Mohammad, 2018) for sentiment analysis, BOLD (Dhamala et al., 2021) for language generation, gendered NLI (Sharma et al., 2020) for natural language inference and Bias in Bios (De-Arteaga et al., 2019) for occupation prediction.

Studies that measure gender bias intrinsically in static word or sentence embeddings measure characteristics of the geometry, such as the proximity between female- and male-related words to stereotypical words, or how embeddings cluster or relate to a gender subspace (Bolukbasi et al., 2016; Caliskan et al., 2017; Gonen and Goldberg, 2019b; Ethayarajh et al., 2019). However, metrics and debiasing methods for static embeddings do not apply directly to contextualized ones. Several studies use sentence templates to adapt to contextual embeddings (May et al., 2019; Kurita et al., 2019; Tan and Celis, 2019). This templated approach is difficult to scale, and lacks the range of representations that a contextual embedding offers. Other work extracts embedding representations of words from natural corpora (Zhao et al., 2019; Guo and Caliskan, 2021; Basta et al., 2019). These studies often adapt the WEAT method (Caliskan et al., 2017), which measures embedding geometry. None measure the effect of the presumably found “bias” on a downstream task.

There is a growing conversation in the field (Barocas et al., 2017; Kate Crawford, 2017; Blodgett et al., 2020; Bommasani et al., 2021) about the importance of articulating the harms of measured bias. In general, extrinsic metrics have clear, interpretable impacts for which harm can be defined. Intrinsic metrics have an unclear effect. Without evidence from a concrete downstream task, a found intrinsic bias is only theoretically harmful. Our work is a step towards understanding whether intrinsic metrics provide valuable insights about bias in a model.

6 Discussion and Conclusions

This study examined whether bias in internal representations is related to extrinsic bias. We designed a new framework in which we debias a model on a downstream task, and measure its intrinsic bias. We found that gender extractability from internal representations, measured by compression rate via MDL probing, reflects bias in a model. Compression was much more reliable than an alternative intrinsic metric for contextualised representations, CEAT. Compression correlated well—to varying degrees—with many extrinsic metrics. We thus encourage NLP practitioners to use compression as an intrinsic indicator for gender bias in NLP models. When comparing two alternative models, a lower compression rate provides confidence in a model's superiority in terms of gender bias. The relative success of compression over CEAT may be because the compression rate was calculated on the same dataset as the extrinsic metrics, whereas CEAT was measured on a different dataset not necessarily aligned with a specific downstream task. The use of a non-task-aligned dataset is a common strategy among other intrinsic metrics (May et al., 2019; Kurita et al., 2019; Basta et al., 2021). Another possible explanation is that compression rate measures a more focused concept, namely the gender information within the internal representations. CEAT measures proximity among embeddings of general terms that may include other social contexts that do not directly relate to gender (e.g. a female term like 'lady' or 'Sarah' contains information about not just gender but class, culture, formality, etc, and it can be hard to isolate just one of these from the rest).

Our results show that when a debiasing method reduces extrinsic metrics but not compression, it indicates that the language model remains biased. When such superficial debiasing occurs, the debiased language model may be reapplied to another task, as in Jin et al. (2021), resulting in unexpected biases and nullifying the supposed debiasing. Our findings suggest that practitioners of NLP should take special care when adopting previously debiased models and inspect them carefully, perhaps using our framework. Our results differ from those of Mendelson and Belinkov (2021a), who found that the debiasing increases bias extractability as measured by compression rate. However, they studied different, non-social biases, that arise from spurious or unintended correlations in training datasets (often called dataset biases). In our case, some debiasing strategies increase intrinsic bias while others decrease it. Future work could investigate why debiasing affects extractability differently for these two types of biases.

Our work also highlighted the importance of the classification layer. Using a debiased objective, such as a balanced dataset, the classification layer can provide significant debiasing. This holds even if the internal representations are biased and the classifier is a single linear layer, as shown in the occupation prediction task. Bias stems in part from internal LM bias and in part from classification bias. Practitioners should focus their efforts on both parts when attempting to debias a model.

We used a broader set of extrinsic metrics than is typically used, and found that the bias metrics behaved differently: some decreased more than others after debiasing, and they correlated differently with compression rate. Debiasing efforts may not be fully understood by testing only a few extrinsic metrics. However, compression as an intrinsic bias metric can indicate meaningful debiasing of internal model representations even when not all metrics are easily measurable, since it correlates well with many extrinsic metrics.

A major limitation of this study is the use of gender as a binary variable, which is trans-exclusive. Cao and Daumé III (2020) made the first steps towards inclusive gender bias evaluation in NLP, revealing that coreference systems fail on gender-inclusive text. Further work is required to adjust our framework to non-binary genders, potentially revealing insights about the poor performance of NLP systems in that area.

Acknowledgements

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 448/20) and by an Azrieli Foundation Early Career Faculty Fellowship. We also thank Kate McCurdy and Andreas Grivas for comments on early drafts, the members of the Technion CS NLP group for their valuable feedback, and the anonymous reviewers for their useful suggestions.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan.

2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2021. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 33(8):3371–3384.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *Computational Linguistics 2021*.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Su Lin Blodgett, Solon Barocas, Hal Daum'e, and H. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *ACL*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

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*.

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

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *arXiv preprint arXiv:2203.13928*.

Maria De-Arteaga, Alexey Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. C. Geyik, K. Kenthapadi, and A. Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

J. Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. 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.

Hila Gonen and Y. Goldberg. 2019a. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL-HLT*.

Hila Gonen and Yoav Goldberg. 2019b. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, 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.

Kate Crawford. 2017. The trouble with bias. keynote at neurips.

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

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

Kweku Kwegyir-Aggrey, Rebecca Santorella, and Sarah M. Brown. 2021. Everything is relative: Understanding fairness with optimal transport. *ArXiv*, abs/2102.10349.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018a. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018b. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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.

Michael Mendelson and Yonatan Belinkov. 2021a. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Mendelson and Yonatan Belinkov. 2021b. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Marnie E Rice and Grant T Harris. 2005. Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r. *Law and human behavior*, 29(5):615–620.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. In *Proceedings of NAACL-HLT*, pages 4187–4195.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2020. Evaluating gender bias in natural language inference. In *NeurIPS 2020 Workshop on Dataset Curation and Security*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

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.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

R. Weischedel, E. Hovy, M. Marcus, and Martha Palmer. 2013. Ontonotes : A large training corpus for enhanced processing. *LDC2013T19, Philadelphia, Penn.: Linguistic Data Consortium.*

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

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Implementation Details

We used RoBERTa in all models (base size, 120M parameters). We use following random seeds in all repeated experiments: 0, 5, 11, 26, 42, 46, 50, 63, 83, 90. Our code was implemented mainly using the Python libraries Pytorch (Paszke et al., 2019), Transformers (Wolf et al., 2020), Sklearn (Pedregosa et al., 2011), and the experiments were logged using Wandb (Biewald, 2020).

A.1 Occupation Classification

We fine-tuned a RoBERTa-base model with a linear classification layer on top. Training was done for 10 epochs at a learning rate of 5e-5, batch size of 64. The input to RoBERTa was the biography tokens, which is limited to the first 128 tokens. The resulting [CLS] token embedding is fed to the classifier to predict the occupation. The probing task involves using the same [CLS] token and training the probing classifier to predict the gender of the person in the biography. The experiments without fine-tuning included either a pre-trained or a previously fine-tuned RoBERTa. We first extracted the pre-trained RoBERTa's embeddings of tokens from the [CLS] and then trained a linear classifier on them. The learning rate was 0.001 and the batch size was 64. We trained the classification layer with pre-trained RoBERTa on 300 epochs, but with fine-tuned RoBERTa, 10 epochs were sufficient. For all training processes, the epoch with the greatest validation accuracy was saved. Fine-tuning took 7 hours on a GeForce RTX 2080 Ti GPU. Bias in Bios contains almost 400k biographies, and we obtain validation (10%) and test set (25%) by splitting with Scikit-learn's (Pedregosa et al., 2011) test_train_split with our random seeds.

A.2 Coreference Resolution

We use the implementation of Xu and Choi (2020), a model that was introduced by Lee et al. (2018b) and has been adopted by many coreference resolution models. Coreference resolution is the process of clustering different mentions in a text that refer to the same real-world entities. The task is solved by detecting mentions through text spans and then predicting for each pair of spans if they represent the same entity. The span representations were extracted with a RoBERTa model, which is fine-tuned throughout the training process, except in the retraining experiment. Fine-tuning took 3 hours on an NVIDIA RTX A6000 GPU. Ontonotes 5.0 has 625k sentences and we use the standard validation and test splits.

A.3 Probing Classifier

We use the MDL probe (Voita and Titov, 2020) implementation by Mendelson and Belinkov (2021b). In all experiments, we use a linear probe and train it with a batch size of 16 and a learning rate of 1e-3. The timestamps used, meaning the accumulating fractions of data that the probe is trained on, are 2.0%, 3.0%, 4.4%, 6.5%, 9.5%, 14.0%, 21.0%, 31.0%, 45.7%, 67.6%, 100%.

A.4 Metrics

A.4.1 Fairness-Based Metrics Implementation

All three statistical fairness metrics measure the difference between two probability distributions, where this difference describes a notion of bias. We calculate Independence and Separation via Kullback–Leibler (KL) divergence, using the AllenNLP implementation (`https://github.com/allenai/allennlp`). We calculate Sufficiency via Wasserstein distance instead, which is motivated by Kwegyir-Aggrey et al. (2021). In this case, we cannot use KL divergence, since there are some classes that do not occur in model predictions for both male and female genders. This causes the probability distributions to not have the same support, and KL divergence is unbounded. Wasserstein distance lacks the requirement for equal support.

A.4.2 Classification Metrics Interpretation in Winobias

Winobias datasets contain pairs of stereotypical and anti-stereotypical sentences. The stereotypes are derived from the US labor statistics (for instance, a profession with a majority of males is stereotypically male). Since coreference resolution is viewed as a clustering problem, it is usually measured via clustering evaluation metrics. Coreference resolution is commonly measured as the average F1 score of these, and the same is true for Winobias. Nevertheless, coreference resolution is accomplished by making a prediction for each pair of mentions, so it can be seen as a classification task. Winobias can be viewed as a simpler task than general coreference resolution, as it contains exactly two mentions of professions and one pronoun, which refers to exactly one profession. Therefore, we reframe it as a classification problem. In a Winobias sentence with two professions x and y, as well as a pronoun p, where p is referring to x, a true positive would be to cluster x and p together, while a false positive would be to cluster y and p together. Our classification metrics are derived based on these definitions. For instance, the TPR gap for profession “teacher”, which is a stereotypical female occupation, is the TPR rate on pro-stereotypical sentences (with a female pronoun) minus the TPR rate on anti-stereotypical sentences (with a male pronoun).

A.4.3 CEAT

The Word Embedding Association Test (WEAT) developed by (Caliskan et al., 2017) is a method for evaluating bias in static word embeddings. The test is defined as follows: given two sets of target words $\mathbb{X}, \mathbb{Y}$ (e.g., ’executive’, ’management’, ’professional’ and ’home’, ’parents’, ’children’) and two sets of attribute words (e.g., male names and female names), and using $\vec{w}$ to represent the word embedding for word w, the effect size is:

$$\text{ES} = \text{mean}_{x\in\mathbb{X}} s(x, \mathbb{A}, \mathbb{B}) - \text{mean}_{y\in\mathbb{Y}} s(y, \mathbb{A}, \mathbb{B})$$

where

$$s(x, \mathbb{A}, \mathbb{B}) = \frac{\text{mean}_{a\in\mathbb{A}} \cos(\vec{\boldsymbol{x}}, \vec{\boldsymbol{a}}) - \text{mean}_{a\in\mathbb{A}} \cos(\vec{\boldsymbol{x}}, \vec{\boldsymbol{b}})}{\text{std-dev}_{w\in\mathbb{X}\bigcup\mathbb{Y}} s(w, \mathbb{A}, \mathbb{B})}$$

In essence, the effect size measures how different are the distances between the embedding vectors of each target group and the attribute groups. Specifically, if $s(x, \mathbb{A}, \mathbb{B}) > 0$, $\vec{\boldsymbol{x}}$ is more similar to attribute words $\mathbb{B}$ and vice versa. For instance, a larger effect size is observed if target words $\mathbb{X}$ are more similar to attribute words $\mathbb{A}$ and target words $\mathbb{Y}$ are more similar to attribute words $\mathbb{B}$. $|ES| > 0.5$ and $|ES| > 0.8$ are considered medium and large effect sizes, respectively (Rice and Harris, 2005). The null hypothesis holds that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words, indicating that there are no biased associations. Statistical significance is defined by the p-value of WEAT, which reflects the probability of observing the effect size under the null hypothesis.

Since a word can take on a great variety of vector representations in a contextual setting, ES varies according to the sentences used to extract word representation. Thus, to adopt WEAT to contextualized representations, the Combined Effect Size (CES) (Guo and Caliskan, 2021) is derived as the distribution of WEAT effect sizes over many possible contextual word representations:

$$\text{CES}(\mathbb{X}, \mathbb{Y}, \mathbb{A}, \mathbb{B}) = \frac{\sum_{i=1}^{N} v_i ES_i}{\sum_{i=1}^{N} v_i}$$

where ES_i denotes the WEAT effect size of the i’th choice of word representations from a large corpus, and v_i is the inverse of the sum of in-sample variance V_i and between-sample variance in the distribution of random-effects. As in Guo and Caliskan (2021), the representation for each word is derived from 10,000 random sentences extracted from a corpus of Reddit comments.

The combined effect size of each of the models is examined on WEAT stimulus 6, which contains target words of career/family and attribute words of male/female names. This was the only one that detected bias on a pre-trained RoBERTa (CES close to 0.5 and $p < 0.05$). The points that we kept in our analysis are those where $p < 0.05$, which make up 90% of the points in occupation prediction and 95% of the points in coreference resolution.

B Full Results

In this section we provide the full results of a RoBERTa model trained on the downstream task. The results for the occupation prediction task after fine-tuning are presented in Table 3 and Table 4 presents the retrained model results. Figure 5 illustrates the correlations between extrinsic metrics and compression rate before and after retraining. Table 5 presents the complete results of the model trained without fine-tuning, meaning that the RoBERTa model is the pretrained version from Liu et al. (2019) and only the classification layer was updated. Subsampling the dataset has significant debiasing effects, which suggests that this debiasing method can achieve low extrinsic bias even when internal bias exists. Table 6 presents the results using a DeBERTa model (He et al., 2020), for the occupation classification task. The trends are similar to those of RoBERTa, with the same metrics showing an increase, no change, or decrease in correlation after re-training, suggesting a general trend in the behavior of these metrics in relation to internal model representations.

Regarding the coreference resolution task, Table 7 displays the results on a finetuned model and Table 8 displays the retraining results. Figure 6 shows the correlations between compression rate and extrinsic metrics before and after the retraining.

	Debiasing Strategy			
Metric	None	Oversampling	Subsampling	Scrubbing
Compression	4.121 ± 1.238	8.522* ± 2.354	3.568 ± 1.516	**1.699*** ± 0.138
Accuracy	**0.861** ± 0.005	0.852* ± 0.004	**0.861** ± 0.003	0.851* ± 0.003
TPR gap (P)	0.763 ± 0.071	0.729 ± 0.067	**0.319*** ± 0.114	0.704* ± 0.068
TPR gap (S)	2.391 ± 0.257	2.145* ± 0.220	**1.598*** ± 0.273	2.019* ± 0.262
FPR gap (P)	0.591 ± 0.052	0.491* ± 0.059	**0.087*** ± 0.094	0.552 ± 0.063
FPR gap (S)	0.075 ± 0.010	0.085* ± 0.011	**0.030*** ± 0.006	0.057* ± 0.007
Precision gap (P)	0.398 ± 0.053	0.327* ± 0.044	**0.166*** ± 0.055	0.347* ± 0.050
Precision gap (S)	0.015 ± 0.001	0.015 ± 0.001	**0.011*** ± 0.001	0.013* ± 0.001
Independence gap (S)	0.009 ± 0.002	0.008 ± 0.002	**0.001*** ± 0.000	0.005* ± 0.001
Separation gap (S)	0.327 ± 0.051	0.305 ± 0.030	**0.204*** ± 0.032	0.296 ± 0.053
Sufficiency gap (S)	9.451 ± 1.945	8.324* ± 1.537	**1.218*** ± 0.330	4.930* ± 0.927

Table 3: Occupation Prediction: Results on a RoBERTa-based model trained over 10 seeds. Significant reduction or increase in a metric ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy is None), is marked with *. The lowest bias score or highest performance metric in each column is marked with **bold**. P = Pearson; S = Sum.

	Debiasing Strategy			
Metric	None	Oversampling	Subsampling	Scrubbing
Compression	4.121 ± 1.238	8.522 ± 2.354	3.568 ± 1.516	1.699 ± 0.138
Accuracy	0.859 ± 0.004	0.856 ± 0.003	0.853 ± 0.003	0.854 ± 0.003
TPR gap (P)	0.777 ± 0.047	0.813* ± 0.040	**0.704*** ± 0.075	0.714* ± 0.068
TPR gap (S)	2.482 ± 0.238	2.593* ± 0.240	2.164* ± 0.284	**1.989*** ± 0.227
FPR gap (P)	0.596 ± 0.041	0.603 ± 0.047	0.602 ± 0.041	**0.536*** ± 0.038
FPR gap (S)	0.073 ± 0.008	0.068* ± 0.007	0.081* ± 0.012	**0.059*** ± 0.005
Precision gap (P)	0.373 ± 0.067	0.356* ± 0.070	0.338* ± 0.054	**0.309*** ± 0.053
Precision gap (S)	0.016 ± 0.002	0.017* ± 0.002	0.015* ± 0.002	**0.014*** ± 0.002
Independence gap (S)	0.009 ± 0.002	0.010* ± 0.002	0.009 ± 0.003	**0.005*** ± 0.001
Separation gap (S)	0.334 ± 0.050	0.328 ± 0.048	0.300* ± 0.049	**0.274*** ± 0.041
Sufficiency gap (S)	9.701 ± 1.305	10.908* ± 1.354	8.370* ± 2.558	**5.239*** ± 0.798

Table 4: Occupation Prediction after retraining: Results on a RoBERTa-based model after retraining of the classification layer with 10 seeds for each pre-trained model. Significant reduction or increase in a metric ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy is None), is marked with *. The lowest bias score or highest performance metric in each column is marked with **bold**. P = Pearson; S = Sum.

	Debiasing Strategy			
Metric	None	Oversampling	Subsampling	Scrubbing
Accuracy	0.824 ± 0.003	0.815* ± 0.005	**0.831*** ± 0.001	0.807* ± 0.003
TPR gap (P)	0.839 ± 0.011	0.443* ± 0.053	**0.158*** ± 0.156	0.814 ± 0.029
TPR gap (S)	3.088 ± 0.192	**1.545*** ± 0.177	1.621* ± 0.088	3.154 ± 0.332
FPR gap (P)	0.598 ± 0.016	0.369* ± 0.029	**0.067*** ± 0.050	0.550* ± 0.012
FPR gap (S)	0.087 ± 0.004	0.041* ± 0.004	**0.027*** ± 0.003	0.112* ± 0.005
Precision gap (P)	0.476 ± 0.027	0.163* ± 0.025	**0.134*** ± 0.065	0.479 ± 0.038
Precision gap (S)	0.017 ± 0.001	0.012* ± 0.001	**0.010*** ± 0.001	0.016* ± 0.002
Independence gap (S)	0.014* ± 0.002	0.001* ± 0.000	**0.000*** ± 0.000	0.022* ± 0.001
Separation gap (S)	0.336* ± 0.044	0.214* ± 0.038	**0.203*** ± 0.024	0.432* ± 0.048
Sufficiency gap (S)	12.019* ± 1.721	2.105* ± 0.576	**1.478*** ± 0.394	13.798* ± 0.966

Table 5: Occupation Prediction: Results on a RoBERTa-based model trained without fine-tuning, over 5 seeds. The compression rate computed on a pre-trained RoBERTa model is 10.122. Significant reduction or increase in a metric ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy is None), is marked with *. The lowest bias score or highest performance metric in each column is marked with **bold**. P = Pearson; S = Sum.

	R^2 Compression		R^2 CEAT	
Metric	Before	After		
TPR gap (P)	0.023	0.120	0.051	0.006
TPR gap (S)	0.000	0.200	0.036	0.098
FPR gap (P)	0.083	0.153	0.121	0.149
FPR gap (S)	0.055	0.013	0.009	0.021
Precision gap (P)	0.065	0.004	0.15	0.025
Precision gap (S)	0.083	0.118	0.027	0.068
Independence gap (S)	0.034	0.084	0.0	0.054
Separation gap (S)	0.000	0.117	0.008	0.009
Sufficiency gap (S)	0.016	0.250	0.046	0.042

Table 6: Results for a DeBERTa model trained on occupation prediction task. Coefficient determination of the regression line taken on the compression rate or CEAT and each extrinsic metric, before and after retraining of the classification layer. P = Pearson; S = Sum. Coefficients are of lower magnitude for DeBERTa than RoBERTa models, but the same trends apply. They largely increase after retraining (save for FPR gap, and a few of the very low magnitude Pearson metrics). The increase after retraining does not apply to CEAT, and the correlations with CEAT are usually lower.

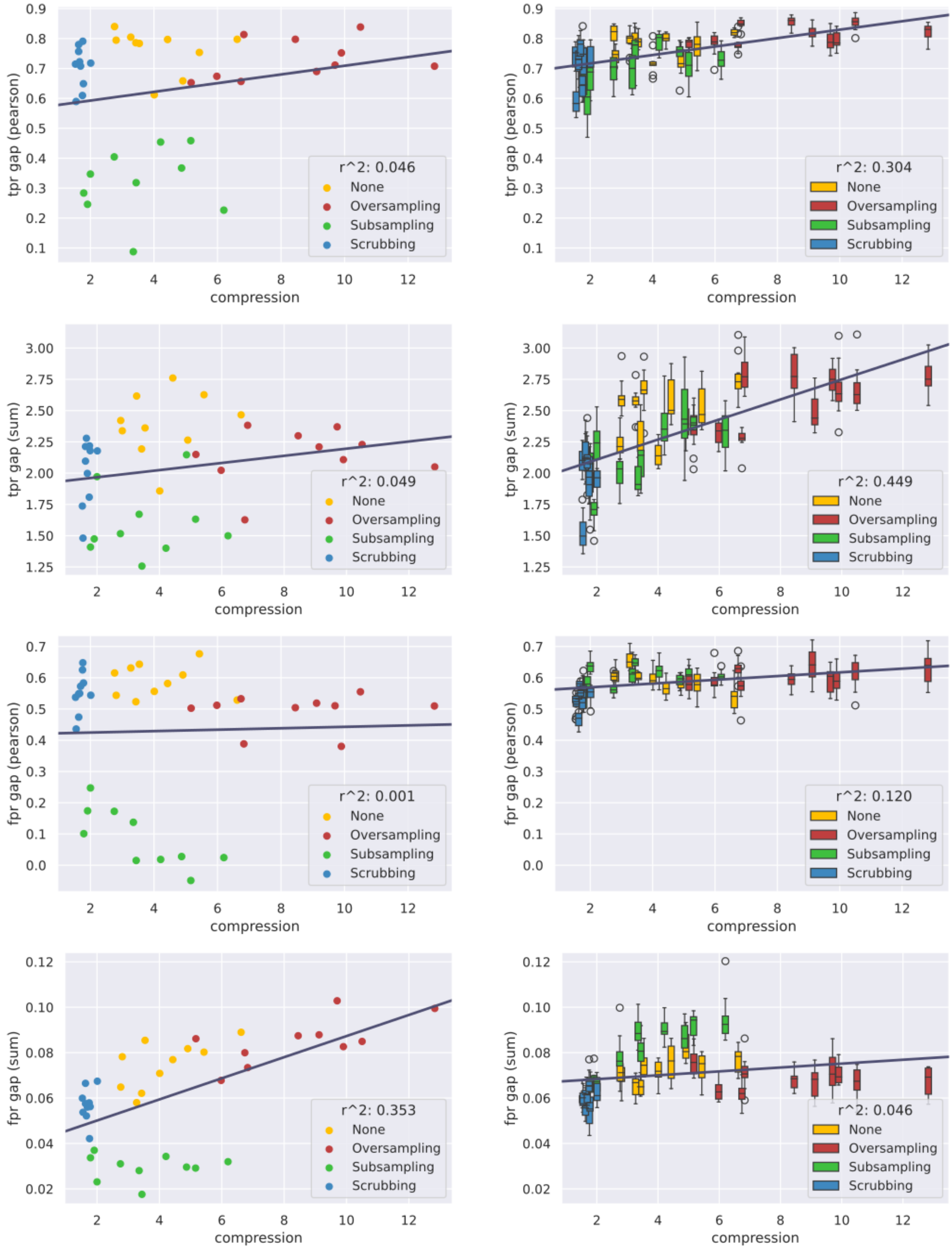


Figure 5: Occupation prediction: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low correlation are discussed in D.1.

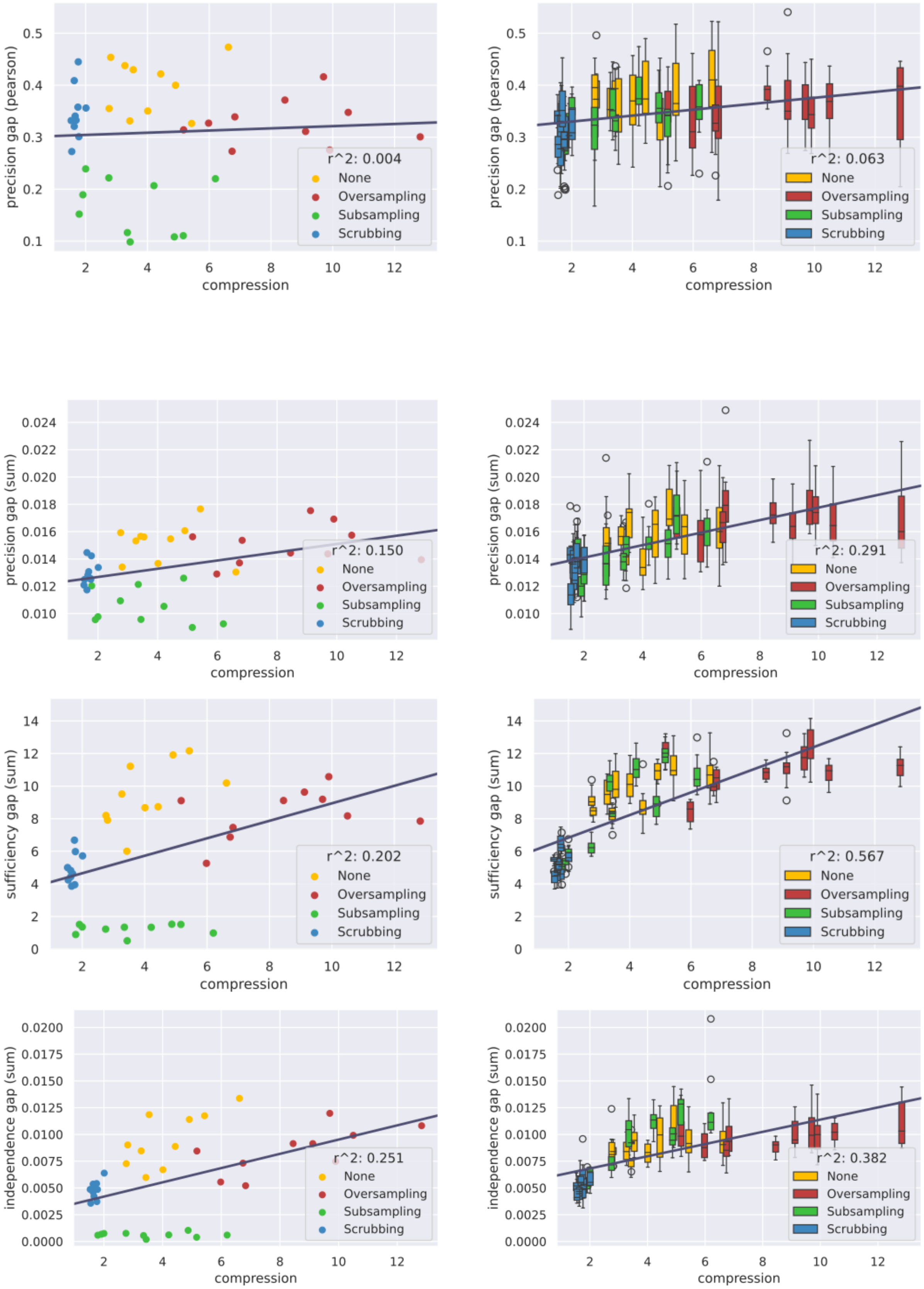


Figure 5: Occupation prediction: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low correlation are discussed in D.1.

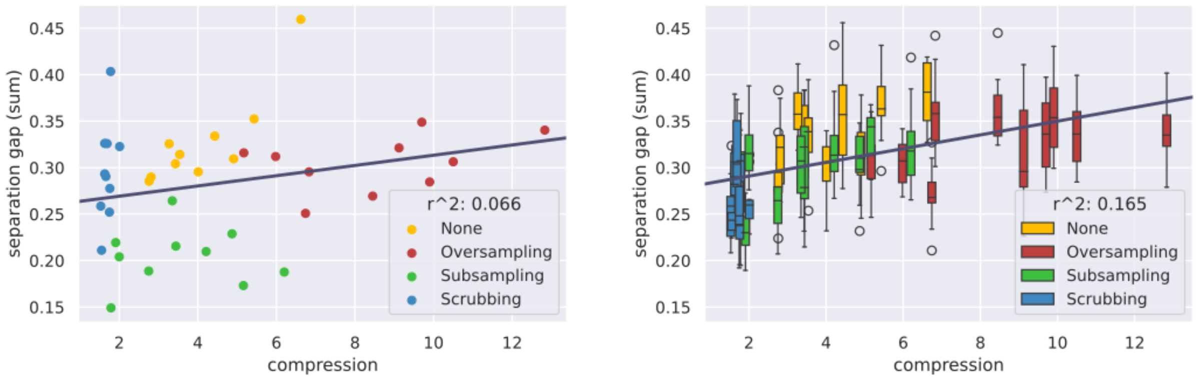


Figure 5: Occupation prediction: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low correlation are discussed in D.1.

	Debiasing Strategy			
Metric	None	Anon	CA	Anon + CA
Compression	1.984 ± 0.101	2.073* ± 0.102	**1.502*** ± 0.075	1.540* ± 0.098
F1 (Ontonotes test)	76.406 ± 0.165	76.538 ± 0.176	77.187* ± 0.071	**77.246*** ± 0.230
F1 diff ($pro - anti$)	6.631 ± 1.013	7.256 ± 0.846	**2.302*** ± 0.466	2.422* ± 0.714
TPR gap (P)	0.654 ± 0.069	0.710* ± 0.047	**0.607** 0.082 ±	0.627 ± 0.100
TPR gap (S)	4.884 ± 0.698	4.870 ± 0.509	2.041* ± 0.228	**2.014*** ± 0.286
FPR gap (P)	0.602 ± 0.036	0.620 ± 0.056	**0.572** ± 0.078	0.629 ± 0.107
FPR gap (S)	0.120 ± 0.015	0.128 ± 0.011	0.050* ± 0.006	**0.049*** ± 0.007
Precision gap (P)	0.654 ± 0.068	0.710* ± 0.048	**0.607** ± 0.083	0.627 ± 0.099
Precision gap (S)	0.061 ± 0.009	0.061 ± 0.006	0.026* ± 0.003	**0.025*** ± 0.004
Independence gap (S)	0.027 ± 0.008	0.025 ± 0.004	**0.004*** ± 0.001	**0.004*** ± 0.001
Separation gap (S)	1.247 ± 0.150	1.344 ± 0.137	**0.537*** ± 0.061	0.557* ± 0.070
Sufficiency gap (S)	8.684 ± 1.883	8.816 ± 1.544	1.673* ± 0.354	**1.557*** ± 0.384

Table 7: Coreference resolution: results on Ontonotes test set and Winobias challenge set. Each model was trained over 10 seeds. * Marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score or highest performance metric in each column is in **bold**. P = Pearson; S = Sum.

	Debiasing Strategy			
Metric	None	Anon	CA	Anon + CA
Compression	1.984 ± 0.065	2.073* ± 0.104	**1.502*** ± 0.081	1.540* ± 0.079
F1 (Ontonotes test)	76.40* ± 0.16	76.48* ± 0.22	76.72* ± 0.15	**76.91*** ± 0.19
F1 diff ($pro - anti$)	6.072 ± 0.789	7.417* ± 1.280	3.674* ± 0.599	**2.858*** ± 0.382
TPR gap (P)	**0.635** ± 0.053	0.688* ± 0.052	0.679* ± 0.062	0.654 ± 0.049
TPR gap (S)	4.561 ± 0.414	5.143* ± 0.713	2.590* ± 0.420	**2.178*** ± 0.201
FPR gap (P)	**0.579** ± 0.046	0.637* ± 0.055	0.620* ± 0.070	0.692* ± 0.075
FPR gap (S)	0.113 ± 0.011	0.126* ± 0.016	0.063* ± 0.010	**0.052*** ± 0.004
Precision gap (P)	**0.636** ± 0.052	0.690* ± 0.052	0.679* ± 0.062	0.652 ± 0.050
Precision gap (S)	0.057 ± 0.005	0.064* ± 0.009	0.032* ± 0.005	**0.027*** ± 0.003
Independence gap (S)	0.022 ± 0.003	0.026* ± 0.006	0.006* ± 0.002	0.004* ± 0.001
Separation gap (S)	1.188 ± 0.114	1.336* ± 0.175	0.670* ± 0.111	**0.594*** ± 0.057
Sufficiency gap (S)	7.350 ± 0.914	8.655* ± 1.726	0.2.401* ± 0.610	**1.653*** ± 0.294

Table 8: Coreference resolution after retraining: results on Ontonotes test set and extrinsic bias metrics on Winobias challenge set. Each model finetuned over 10 seeds and re-trained over 5 seeds. * Marks significant reduction or increase in bias ($p < 0.05$ on Pitman's permutation test), compared to the non-debiased model (debiasing strategy None). The lowest bias score or highest performance metric in each column is in **bold**. P = Pearson; S = Sum.

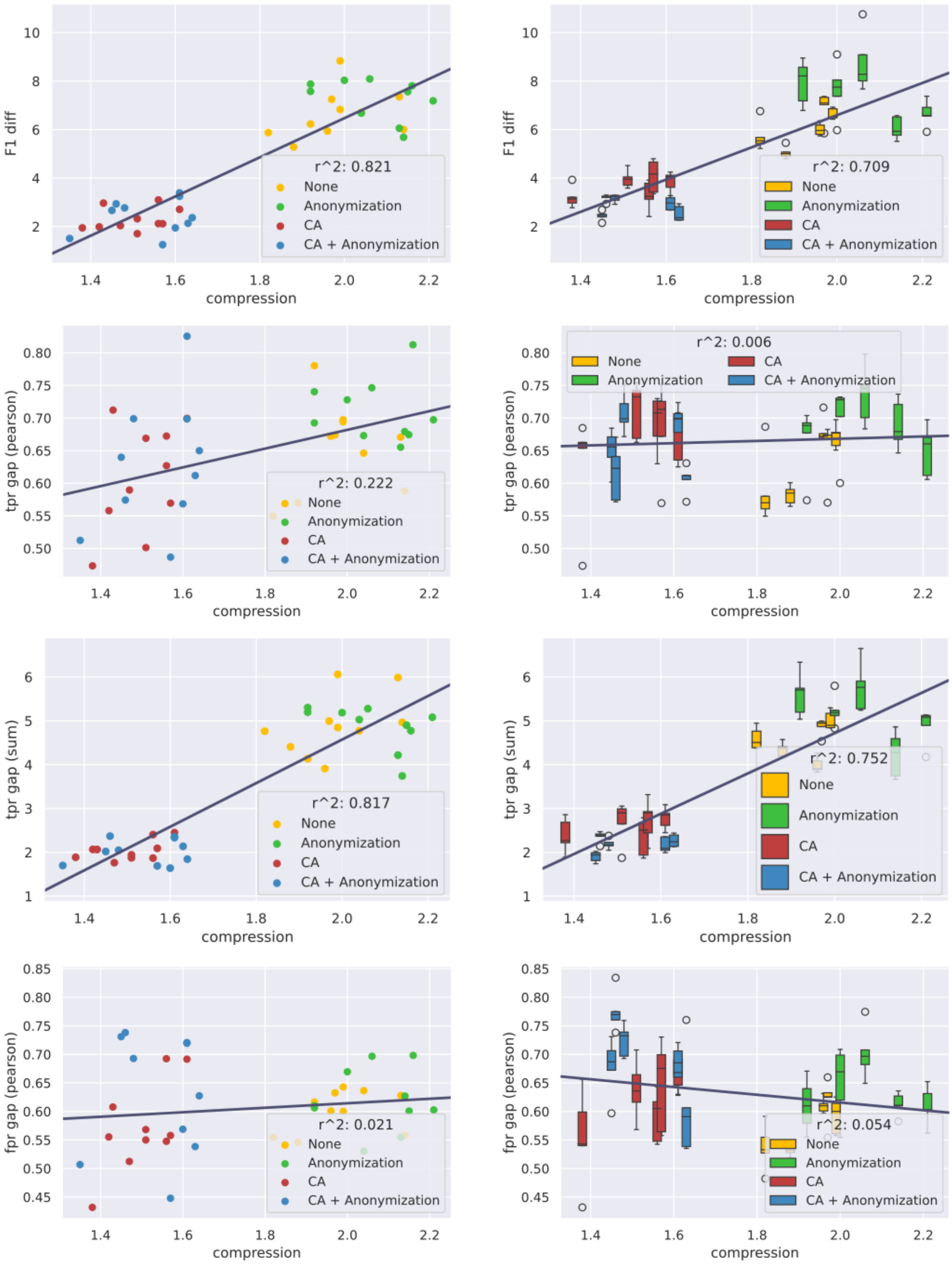


Figure 6: Coreference resolution: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low and no correlation with the Pearson metrics are discussed in D.2.

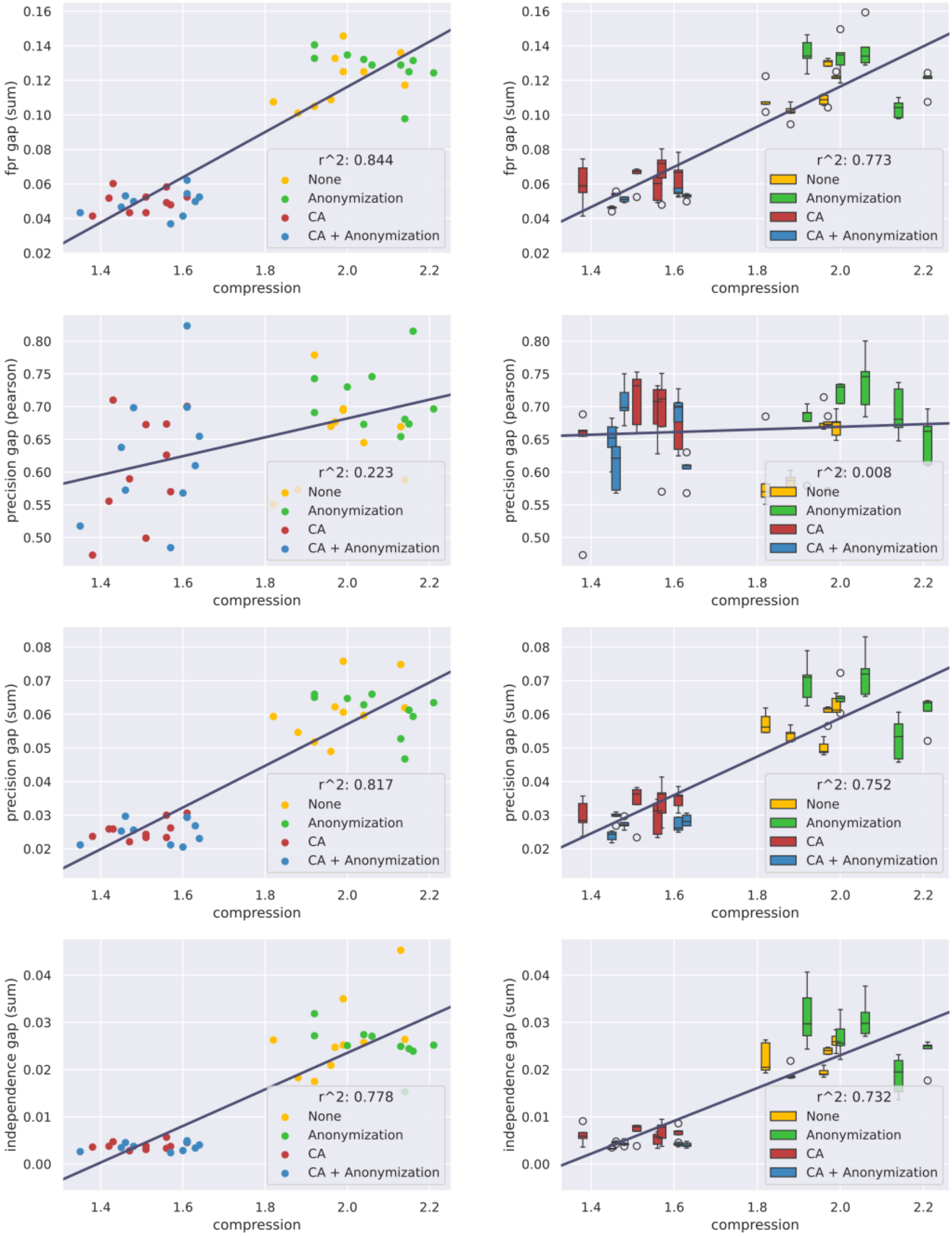


Figure 6: Coreference resolution: Before (left) and after (right) plots of compression rate versus and extrinsic metric.

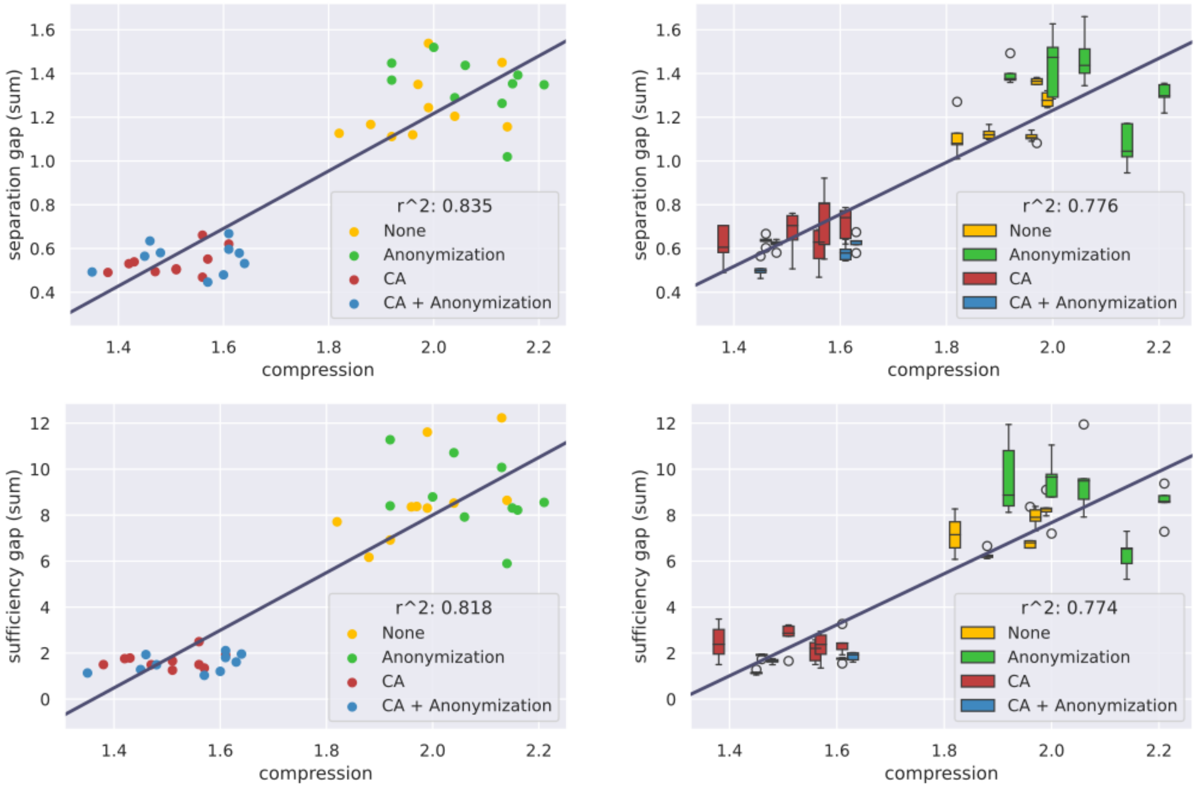


Figure 6: Coreference resolution: Before (left) and after (right) plots of compression rate versus and extrinsic metric. Cases of low and no correlation with the Pearson metrics are discussed in D.2.

Female Words	**Male Words**
husband, women,	chief, companies
gender, listed,	computer,
practices, nurse,	applications,
specializes,	md, accepts,
children,	known, doctors,
ba, child,	npi, sports,
reading, families,	philosoph',
location, place,	problems, rating,
affiliated, family,	no, systems,
experiences,	theory, practicing,
spanish,	software,
love, justice	security, major

Table 9: Top 20 significant words used to predict gender on all biographies, as obtained from a logistic regression model trained on predicting the gender of a person described in a biography. The words are sorted by importance.

Female Words	**Male Words**
husband , women,	holds , emergency,
midwife , providing	vanderbilt, forces,
book , includes,	registered, mental,
joining, faculty	assistant, president

Table 10: Top 8 words used to predict gender of female and male nurses, as obtained from a logistic regression model trained on predicting the gender of a person described in a biography. The words are sorted by importance.

C Why is scrubbing not as effective as subsampling?

The debiasing method of subsampling significantly reduced external biases in the occupation prediction task. Although compression rates show that scrubbing reduced more gender information, subsampling outperforms it as a debiasing method. We find that in spite of the scrubbing, a probe is able to correctly identify the gender from an internal representation with 68.8% accuracy compared to 90.7% on the original, non-scrubbed data. This means that although the scrubbing process reduces extrinsic bias significantly, gender information is still embedded in the [CLS] token embeddings.

To investigate the source of gender information after scrubbing, we use logistic regression (LR) model to predict the gender from the Bag-of-Words of the scrubbed biographies. We perform an iterative process for automatic extra scrubbing: in each iteration we (1) train a LR model for gender prediction (2) scrub the n most significant words for each gender according to the LR weights. The most relevant words among 5 seeds of training with n=10 words scrubbed per iteration are displayed in Table 9. The model learns indirect correlations to gender in the absence of explicit gendered words. Because the significant words are related to male- or female-dominated professions, we conducted the process on a specific profession. Table 10 presents the most significant words for biographies of nurses. There are differences in wording even between females and males in the same profession. The results of this study are in line with the results of other studies that have been conducted on the way biographies are written for men and women (Wagner et al., 2016; Sun and Peng, 2021).

Subsampling is therefore more effective even when gender information is present since it prevents the model from learning correlations between gender information and a profession whereas scrubbing only attempts to remove gender indicators without removing correlations. On the other hand, it is possible that oversampling is less effective for debiasing since seeing more non-unique examples an unrepresented group encourages learning correlations.

D A closer look into no-correlation cases

D.1 Occupation Prediction

Although compression has the ability to identify bias in most cases, some metrics still show little or no correlation with compression rate. These results suggest that gender information comprises only one facet of embedded bias in the representations. Other factors that may influence these metrics are not considered or measured, such as the connection between a name and a profession.

For example, as can be see in Tables 3 and 4, LMs finetuned on subsampled data have the largest FPR gaps after retraining, despite being the least biased before retraining, while those finetuned on oversampled data have the next-to-lowest FPR gaps after retraining. The information encoded in the internal representations may have been encoded in a manner that allowed the classification layer to exhibit a smaller FPR gap when trained on a balanced dataset. However, when the classification layer was retrained on biased training data, it used the same features to make biased predictions.

D.2 Coreference Resolution

The cases where there is no correlation between our intrinsic metric and an extrinsic metric are the cases where the metric is based on Pearson correlation. Unlike occupation prediction, coreference resolution seems to exhibit no correlation between those metrics and compression rate. These metrics are computed as the Pearson correlation between a performance gap for a specific profession and the percentage of women in that profession, however the percentages are computed differently in each task: in occupation prediction, the percentages are computed from the train set, focusing on the representation each gender has in the data. In Winobias, the percentages are taken from the US labor statistics, and are unrelated to the training dataset statistics. We note that the two statistics can be different - the real-world representation of women in a profession does not have to be equal to their representation in written text (Suresh and Guttag, 2021). We thus decided to test what happens if we change the statistics used in Winobias to dataset statistics, but Ontonotes 5.0 has very little representation to each profession and the statistics extracted from it would not be reliable. We thus took a different approach and computed the Pearson correlations for occupation prediction with real world statistics instead of dataset statistics. To do this, we mapped the professions appearing in this dataset to professions from the US labor statistics, and dropped those who could no be mapped (6 out of 29 of the professions which is 21.4%). We then repeated all experiments on the Pearson metrics using these statistics. Figure 7 shows the results. Correlations are very different when computed with respect to real-world statistics. TPR-gap has no correlation at all although it had with training data statistics, the correlation for FPR-gap after retraining exists but is negative, and the correlation with precision-gap does not exist after retraining. We thus conclude that the Pearson metrics are less reliable as they are heavily dependent on the statistics with respect to which they are calculated.

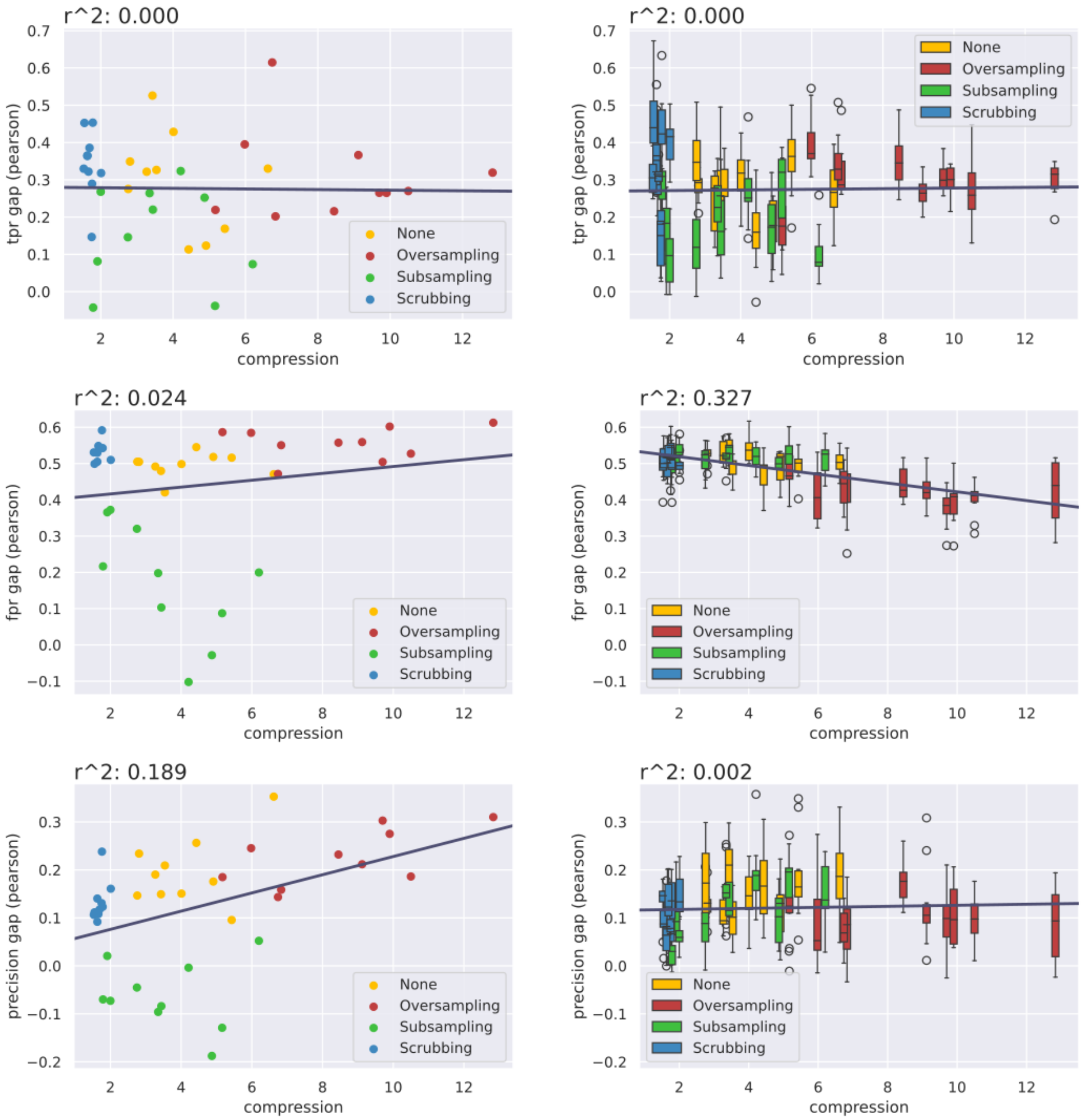


Figure 7: Occupation prediction: Before (left) and after (right) plots of compression rate versus Pearson metrics as computed from real-world statistics (as opposed to statistics derived from the training dataset). This shows the unrealiability of using real world statistics to draw conclusions, as they may not be reflected in the data.