

Modular and On-demand Bias Mitigation with Attribute-Removal Subnetworks

Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar,
Markus Schedl, Navid Rekabsaz

Johannes Kepler University Linz, Austria
Linz Institute of Technology, AI Lab

{first_name.family_name}@jku.at

Abstract

Societal biases are reflected in large pre-trained language models and their fine-tuned versions on downstream tasks. Common in-processing bias mitigation approaches, such as adversarial training and mutual information removal, introduce additional optimization criteria, and update the model to reach a new debiased state. However, in practice, end-users and practitioners might prefer to switch back to the original model, or apply debiasing only on a specific subset of protected attributes. To enable this, we propose a novel modular bias mitigation approach, consisting of stand-alone highly sparse debiasing subnetworks, where each debiasing module can be integrated into the core model on-demand at inference time. Our approach draws from the concept of *diff* pruning, and proposes a novel training regime adaptable to various representation disentanglement optimizations. We conduct experiments on three classification tasks with gender, race, and age as protected attributes. The results show that our modular approach, while maintaining task performance, improves (or at least remains on-par with) the effectiveness of bias mitigation in comparison with baseline finetuning. Particularly on a two-attribute dataset, our approach with separately learned debiasing subnetworks shows effective utilization of either or both the subnetworks for selective bias mitigation.

1 Introduction

A large body of research evidences the existence of societal biases and stereotypes in pre-trained language models (PLMs) (Zhao et al., 2019; Sheng et al., 2019; Rekabsaz et al., 2021), and their potential harms when used in down-stream tasks (Blodgett et al., 2020; De-Arteaga et al., 2019; Rekabsaz and Schedl, 2020; Stanovsky et al., 2019). Common in-processing approaches to bias mitigation update a model’s (typically all) parameters to satisfy specific attribute erasure criteria through optimization methods such as adversarial train-

ing (Elazar and Goldberg, 2018; Rekabsaz et al., 2021), and mutual information reduction (Colombo et al., 2021). These methods are shown to be effective in reducing the footprint of protected attributes (e.g., gender, race, etc.) in the resulting model.

However when using such debiasing models in practice and in specific use-cases, system designers or end-users might still prefer to instead use the original model, or a debiased variation in respect to a particular subset of protected attributes. This can be due to various reasons such as the nature of a given input, preference of an individual end-user, or fairness-utility trade-off considerations. For instance, while a bias-aware model should indeed be agnostic to genders when the input is about gender-neutral occupations (such as *nurse* or *CEO*), certain topics like *pregnancy* may specifically require gender information for a correct model decision.¹ Also as shown in previous studies (Zerveas et al., 2022; Biega et al., 2018; Rekabsaz et al., 2021), since improving fairness on specific tasks may come with the cost of performance degradation, it is necessary to provide on-demand control over whether to impose fairness/debiasing criteria. Using existing approaches, this would require maintaining and deploying multiple large parallel models for every protected attribute, resulting in overly complex and resource-heavy pipelines and increased latency.

To address this, we introduce a novel modular bias mitigation approach using sparse weight-difference networks. In our approach, the required changes in a model’s parameters for erasing a bias attribute are stored in a decoupled subnetwork, trained simultaneously by a debiasing and a sparsification objective. At inference time, adding each debiasing module to the core model results in delivering debiasing qualities to a model’s prediction in respect to the corresponding protected attribute. Our approach extends the principle idea

¹See also the discussion in Krieg et al. (2023) about the need to separate bias-sensitive queries from “normal” ones.

of *diff pruning* (Guo et al., 2021) introduced for parameter-efficient task training to bias mitigation, by viewing the objective of erasing a protected attribute as a stand-alone *diff* module. This module replaces fine-tuning by training only a small set of parameters added to the corresponding PLM’s parameters, to deliver bias mitigation of a specific protected attribute. We further propose a novel procedure to train such debiasing subnetworks separately, and to selectively add an arbitrary set of them to the core model at inference time (§3).

Our approach can be applied to any debiasing and representation disentanglement method, provided that its objective has separate learning signals for the task and each protected attribute. In comparison with adapter networks (Rebuffi et al., 2017; Houlsby et al., 2019), as shown by Guo et al. (2021) and also evidenced in our experiments, even more parameter-efficiency can be provided. Additionally, since our approach extends the base model with *diff* subnetworks, the resulting model is expected to perform (at least) as good as the fine-tuning variation, avoiding possible performance degradations. The modularity of our approach supports separating the process of developing debiasing solutions for a task from using them, such that stand-alone debiasing modules can be created and shared, and later be utilized in a final system on-demand.

We evaluate our approach on three bias mitigation tasks: occupation prediction from biographies involving gender (De-Arteaga et al., 2019), hate speech detection with dialect-based race as sensitive attribute (Founta et al., 2018); and mention prediction in tweets with two attributes of gender and age of the authors (Pardo et al., 2016). The last dataset particularly enables the study of combining independently trained debiasing modules (details in §4). The evaluation results show that our approach, due to learning the subnetworks specialized on the narrow functionality of debiasing an attribute, provides on par or better debiasing performance in comparison with strong baselines. Additionally, we observe that on the mention detection task, learning the debiasing subnetworks post-hoc to model training provides effective results when combining the two (independently trained) subnetworks at inference time. Remarkably, these results are achieved with debiasing subnetworks of maximum 1% size of the core model (BERT-Base), in some cases (e.g., for gender attribute) even only 0.01% (details in §5).

2 Related Work

2.1 Parameter-efficient and Modular Training

The concept of subnetworks is grounded in the lottery ticket hypothesis (Frankle and Carbin, 2019), stating that in deep neural networks, one can find many sparse subnetworks with a capacity comparable to that of the base network. Zhou et al. (2019) show that spotting such subnetworks through binary masks can in fact be seen as a form of model training. Zhao et al. (2020) further consider the magnitude of the parameters for pruning and filter out the ones lower than a specific threshold. Guo et al. (2021) use L_0 -regularization (Louizos et al., 2018) to reduce the number of active neurons. Hu et al. (2021) propose low-rank adaptation via rank decomposition matrices. These methods are extended by structural pruning approaches with additional architectural constraints, commonly led to a higher parameter-efficiency (Lagunas et al., 2021; Jaszczerzak et al., 2021).

Recent studies exploit subnetworks in the context of multi-task learning. Wortsman et al. (2020) isolate the learning signal of each task in a separate masking subnetwork, while Ben-Zaken et al. (2021) only finetune the bias weights. Xu et al. (2021) learn a subset of parameters via masking out the gradients of other parameters during backward-pass. Guo et al. (2021) suggest learning a sparse *diff* subnetwork for each task, whose parameter values are added to the corresponding parameters of the base network. An alternative architecture are adapter networks, first introduced in the context of multi-task learning (Rebuffi et al., 2017; Houlsby et al., 2019; Stickland and Murray, 2019), and are then extended in respect to their parameter efficiency (Rücklé et al., 2021; Han et al., 2021a), architectural variations (Mahabadi et al., 2021), and transfer learning capacity (Pfeiffer et al., 2021). As stated by (Sung et al., 2021), adapters (in their original form) slightly increase the inference cost of a model in comparison with the original form or pruning-based variations. Our proposed approach contributes to this line of research by extending this concept to on-demand bias mitigation.

2.2 Fairness & Bias Mitigation in NLP

Several studies explore methods for debiasing PLMs, such as linearly projecting embeddings into the space with minimum correlations to protected features (Ravfogel et al., 2020; Kaneko and Bollegala, 2021; Bolukbasi et al., 2016), utilizing a

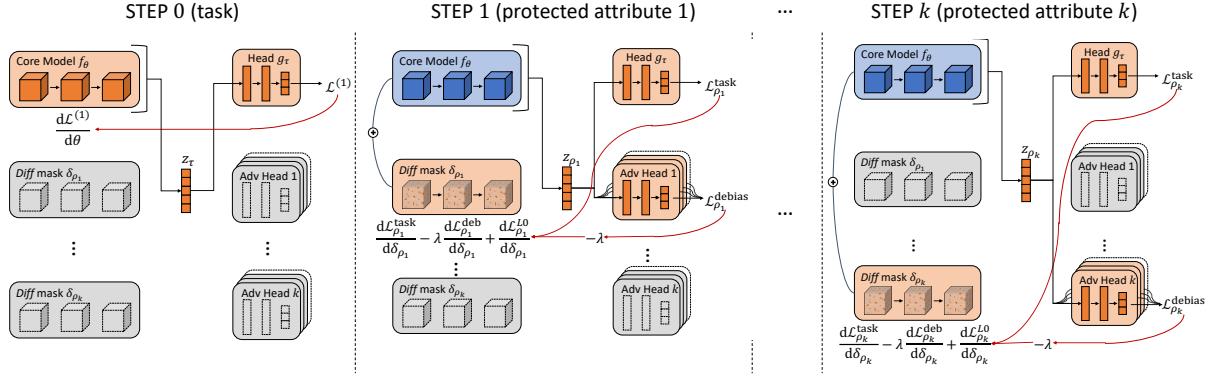


Figure 1: Training procedure of MoDDIFFY-PAR in one batch with adversarial bias removal for k protected attributes. The color orange indicates the trainable parameters during each step, while blue shows the frozen ones. The core model learns the task in Step 0, and the subnetworks — each responsible for debiasing a protected attribute — are trained in the next steps. In MoDDIFFY-POST training regime, after finetuning the core model in Step 0, the task head parameters (together with the ones of the core model) remain frozen in the next steps.

distribution alignment loss (Guo et al., 2022), or penalizing bias by utilizing the encoded information in models (Schick et al., 2021). Adversarial training, originally introduced in domain adaptation (Ganin et al., 2016; Ganin and Lempitsky, 2015) is utilized in the context of fair representation learning (Xie et al., 2017; Madras et al., 2018), and later to erase demographic data from text classifiers (Elazar and Goldberg, 2018; Barrett et al., 2019; Han et al., 2021b; Wang et al., 2021), information retrieval models (Rekabsaz et al., 2021), and recommendation systems (Ganhör et al., 2022). Mutual information removal is an alternative approach, which minimizes the approximate upper bound of the common information between the task and protected attributes (Cheng et al., 2020; Colombo et al., 2021). Our work utilizes these optimizations to learn a novel modular on-demand bias mitigation approach.

Few recent studies explore parameter-efficient training for bias mitigation. Lauscher et al. (2021) approach debiasing PLMs using a stack of adapters. While shown effective in practice, the adapters in the higher levels inherently depend on the ones in the lower levels and cannot be learned nor utilized stand-alone. Zhang et al. (2021) approach bias mitigation with binary masks applied to the base network. More recently and in the context of removing spurious shortcuts in natural language understanding datasets, Meissner et al. (2022) train sparse binary masks on a finetuned model. Our work extends this line of research by encapsulating concept erasure modules in separate *diff* subnetworks for each protected attribute, and selectively

applying them to the base model at inference time.

3 Modular Debiasing with *Diff* Subnets

We start with defining the general approach to model bias mitigation. We consider an arbitrary PLM denoted by f_θ with the set of parameters θ . The model learns the task τ using the loss function \mathcal{L}_τ . The predictions of f_θ might be sensitive to the variations in any of the k protected attributes of the set $P = \{\rho_1, \dots, \rho_k\}$. The bias mitigation objective is to make the model invariant to these variations while maintaining the effectiveness on the task, approached by defining the debiasing loss \mathcal{L}_{ρ_i} for the protected attribute ρ_i . We discuss two realizations of this loss function in Section 3.2. A debiased model in respect to attributes P is achieved by training on the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_\tau + \mathcal{L}_{\rho_1} + \dots + \mathcal{L}_{\rho_k} \quad (1)$$

Using \mathcal{L}_{total} , one can finetune all parameters of the model (Elazar and Goldberg, 2018; Rekabsaz et al., 2021), or utilize any parameter-efficient training such as adapters (Lauscher et al., 2021), *diff* pruning (Guo et al., 2021), or binary masks (Zhang et al., 2021). We use some of these methods as baselines, explained in the following sections. In the remainder of this section, we introduce our *Modular Debiasing with Diff Subnetworks* (MoDDIFFY), explain its training and inference procedure, and describe two debiasing optimization methods.

3.1 MoDDIFFY

We aim to encapsulate the debiasing functionality of the protected attribute ρ_i , provided by the signal

of the corresponding loss \mathcal{L}_{ρ_i} , into the sparse *diff* subnetwork characterized by the set of parameters δ_{ρ_i} . Each parameter in δ_{ρ_i} corresponds to a parameter in θ , such that adding δ_{ρ_i} to the corresponding parameters in θ results in debiasing the model f .

To learn the model, let us consider a training data item in form of $\langle x, y_\tau, y_{\rho_1}, \dots, y_{\rho_k} \rangle$, where x is the input, y_τ the task label, and y_{ρ_i} denotes the label of the corresponding protected attribute ρ_i . Figure 1 depicts the MODDIFFY approach with adversarial bias mitigation optimization. We first optimize for the task by encoding x into the vector z_τ using $f(\cdot; \theta)$. An arbitrary decoder network denoted as g_τ uses z_τ to predict the task output, and task loss is calculated using cross entropy (CE), as formulated below:

$$z_\tau = f(x; \theta), \quad \mathcal{L}_{total}^{(0)} = \text{CE}(g_\tau(z_\tau), y_\tau) \quad (2)$$

This loss updates θ as well as the parameters of g_τ . While in our experiments we opt for fully finetuning f_θ , in practice the model can be trained using any parameter-efficient method.

Next, we iterate over the protected attributes, and in each step learn the corresponding debiasing *diff* subnetwork. Specifically in step i dedicated to the protected attribute ρ_i , we learn the set of sparse additional parameters δ_{ρ_i} such that by being added to θ , the resulting $f(\cdot; \theta + \delta_{\rho_i})$ model is debiased. Learning δ_{ρ_i} is characterized by three loss functions. The first loss, $\mathcal{L}_{\rho_i}^{\text{task}}$ maintains the task performance when the task output is predicted from the altered encoder formulated below:

$$z_{\rho_i} = f(x; \theta + \delta_{\rho_i}), \quad \mathcal{L}_{\rho_i}^{\text{task}} = \text{CE}(g_\tau(z_{\rho_i}), y_\tau) \quad (3)$$

The second is the debiasing loss $\mathcal{L}_{\rho_i}^{\text{debias}}$, defined based on z_{ρ_i} and the label of the corresponding protected attribute y_{ρ_i} . We discuss adversarial bias removal and mutual information reduction as two possible realization of this representation disentanglement loss in Section 3.2. The third loss imposes the sparsity constraint, defined as the L_0 regularization of δ_{ρ_i} . The L_0 loss aims to reduce the number of non-zero parameters, namely the term $\sum_{j=1}^{|\delta_{\rho_i}|} \mathbb{1}\{\delta_{\rho_i,j} \neq 0\}$, and is realized with the differentiable approximation proposed by Louizos et al. (2018). Following Guo et al. (2021), δ_{ρ_i} is decomposed into the element-wise multiplication of two sets of parameters: $\delta_{\rho_i} = \mathbf{m}_{\rho_i} \odot \mathbf{w}_{\rho_i}$. The parameter set \mathbf{w}_{ρ_i} stores the magnitude changes (*diff* values), and \mathbf{m}_{ρ_i} learns to mask out the parameters.

\mathbf{m}_{ρ_i} is characterized by the hard concrete distribution (Guo et al., 2021) with $(\log \alpha_{\rho_i}, 1)$ parameters, and $\gamma < 0$ and $\zeta > 1$ hyperparameters. This results in the following formulation:

$$\mathcal{L}_{\rho_i}^{L_0} = \sum_{j=1}^{|\delta_{\rho_i}|} \sigma \left(\log \alpha_{\rho_i,j} - \log \left(-\frac{\gamma}{\zeta} \right) \right) \quad (4)$$

where σ denotes the sigmoid function, and as in Guo et al. (2021) β is set to 1. The masking network can be simply reduced by assigning a mask to a group of parameters (such as a weight matrix, or a layer) instead of each individual one.

Putting all together, the objective of MODDIFFY at step $i > 0$ is defined as:

$$\mathcal{L}_{total}^{(i)} = \mathcal{L}_{\rho_i}^{\text{task}} + \mathcal{L}_{\rho_i}^{\text{debias}} + \mathcal{L}_{\rho_i}^{L_0} \quad (5)$$

We should note that while each subnetwork is trained independently from the others, the output embeddings z_{ρ_i} are passed to the same decoder network g_τ . This design choice forces the learned embeddings to remain in the same embedding space, making it possible to add multiple (independently trained) subnetworks together to the core model. Another aspect is that the sparsity rate of each resulting subnetwork is not fixed and may vary due to the factors such as the hyperparameter setting, and the extent of encoded information content. To achieve a fixed sparsity rate, we further apply magnitude pruning by only keeping a fixed portion of the parameters with the largest absolute values, and finetuning the resulting parameters with the task and debiasing loss terms.

We conduct optimization with \mathcal{L}_{total} under two training regimes. In the first one referred to as MODDIFFY-PAR, we repeat the mentioned training steps for each training batch. The second approach referred to as MODDIFFY-POST trains debiasing subnetworks post-hoc to training the core model. While MODDIFFY-PAR accommodates for a higher flexibility in optimization by training the networks in parallel, MODDIFFY-POST provides the practical benefits of learning various debiasing solutions for an already trained core model.

Finally at *inference time*, one can use the core model in its original form $f(x; \theta)$, or in combination with any of the debiasing subnetworks in the form of $f(x; \theta + \delta_{\rho_i})$. Our approach also enables simultaneously debiasing all or any subset of the protected attributes by adding their corresponding subnetworks to the core network, for instance

as in $f(x; \theta + \delta_{\rho_1} + \dots + \delta_{\rho_k})$. We should note that, although these subnetworks reside in the same distributional space, they might affect each others functionality, depending on the inherent nature of the biases as well as the possible correlations between the protected attributes. We will examine this case in the next sections on a dataset with two protected attributes of gender and age.

3.2 Bias Mitigation Objectives

We explain two bias mitigation optimization methods used in MODDIFFY in the following.

Adversarial Bias Removal This method first defines a new classification head h_{ρ_i} for each protected attribute ρ_i . This head receives z_{ρ_i} as input, predicts the corresponding protected attribute, and calculates the cross entropy loss function $\mathcal{L}_{\rho_i}^{\text{debias}}$. This loss needs to remove the information of ρ_i from f but train h_{ρ_i} such that it can effectively predict the protected attribute. This optimization forms the min-max game: $\mathcal{L}_{\rho_i}^{\text{debias}} = \min_f \max_{h_{\rho_i}} \text{CE}(h_{\rho_i}(z_{\rho_i}), y_{\rho_i})$. A common approach to turn this loss into a minimization problem is by using gradient reversal layer (GRL) (Ganin and Lempitsky, 2015) added before the debiasing heads. GRL multiplies the gradient of $\mathcal{L}_{\rho_i}^{\text{debias}}$ with a factor of $-\lambda_i$, and thereby simplifies the learning process to a standard gradient-based optimization, formulated below:

$$\mathcal{L}_{\rho_i}^{\text{debias}} = \min_{f, h_{\rho_i}} \text{CE}(h_{\rho_i}(z_{\rho_i}), y_{\rho_i})$$

Mutual Information (MI) Reduction This approach represents a family of algorithms that aims to remove the mutual information of the encoded embeddings of the task and protected attributes. Maximum Mean Discrepancy (MMD), first introduced in the context of domain adaptation (Gretton et al., 2012; Tzeng et al., 2014), offers a realization of MI reduction by minimizing the ability to separate the subsets belonging to two protected attributes. In particular, given a set of data points X split into two subsets $X_{\rho_i}^A$ and $X_{\rho_i}^B$ according to the values of the (binary) protected attribute ρ_i , MMD minimizes the distance between the encoded embeddings of the subgroups with the following loss formulation:

$$\mathcal{L}_{\rho_i}^{\text{debias}} = \left(\frac{\sum_{x^A \in X_{\rho_i}^A} \phi(f(x^A))}{|X_{\rho_i}^A|} - \frac{\sum_{x^B \in X_{\rho_i}^B} \phi(f(x^B))}{|X_{\rho_i}^B|} \right)^2$$

where ϕ is the feature map kernel defined as a linear combination of multiple Gaussian kernels.

4 Experiment Design

Datasets We evaluate our approach on three datasets on the tasks of occupation prediction, hate speech detection, and mention prediction; involving protected attributes of gender, age, and race dialect. The first dataset is **BIOS** (De-Arteaga et al., 2019) which contains short biographies used to predict a person’s job, where the name and any indication of the person’s gender (such as pronouns) in the biography are omitted. The BIOS dataset contains around 430K data points with 28 occupations, and two protected attribute classes (female/male). The second dataset is **FDCL18** (Founta et al., 2018) for hate speech detection, containing a set of tweets each classified as hateful, abusive, spam, or none. As discussed in Xia et al. (2020), hate speech might have a strong correlation with dialect-based racial bias. Following previous studies (Sap et al., 2019; Ravfogel et al., 2020), we assign race dialect labels of *African American* and *White American* to FDCL18 using the probabilistic model developed by Blodgett et al. (2016), resulting in the dataset of approximately 62K data points. The third dataset is **PAN16** (Rangel et al., 2016) containing a set of tweets accompanied with the labels of gender and age of the authors. The task’s objective is to predict mentions (whether another user is mentioned in a tweet). PAN16 provides approximately 200K data points with binary task classes (*mention*, *no mention*), as well as two gender labels and five age groups. Further details on the three datasets are provided in Appendix A.

Models and Baselines We conduct the experiments on the following models and baselines. **FINETUNE**: finetuning all parameters of the PLM on the task without any bias mitigation objective. **FINETUNE-DEBIAS**: the same model as FINETUNE but with the bias mitigation objective. **ADAPTER**: learning the task with an adapter network while the rest of the PLM’s parameters are kept frozen. **ADAPTER-DEBIAS**: the same model as ADAPTER but the adapter is trained on both task and bias mitigation objectives. **DIFFPRUN**: using a *diff* network to learn the task while PLM parameters remain unchanged. **DIFFPRUN-DEBIAS**: same as DIFFPRUN but the *diff* network is trained on both task and bias mitigation objectives. **MODDIFFY-POST**: our introduced post-hoc approach

Model	BIOS (gender)				FCDL18 (race-dialect)			
	Adversarial		MI Reduction		Adversarial		MI Reduction	
	Task↑	Probe↓	Task↑	Probe↓	Task↑	Probe↓	Task↑	Probe↓
FINETUNE	84.1 _{0.3}	67.2 _{0.7}	84.1 _{0.3}	67.2 _{0.7}	82.0 _{0.4}	93.0 _{0.4}	82.0 _{0.4}	93.0 _{0.4}
ADAPTER	84.4 _{0.1}	65.9 _{0.1}	84.4 _{0.1}	65.9 _{0.1}	80.9 _{0.1}	82.1 _{4.1}	80.9 _{0.1}	82.1 _{4.1}
FINETUNE-DEBIAS	84.2 _{0.1}	56.9 _{0.9}	84.1 _{0.7}	61.7 _{0.8}	81.9 _{0.7}	84.5 _{3.9}	81.6 _{0.4}	87.3 _{0.5}
ADAPTER-DEBIAS	84.6 _{0.2}	60.8 _{0.4}	84.4 _{0.0}	65.6 _{0.2}	80.5 _{0.7}	65.6 _{0.1} ♣	80.1 _{0.4}	82.1 _{1.3} ♣
DIFFPRUN	84.6 _{0.1}	68.9 _{0.2}	84.6 _{0.1}	68.9 _{0.2}	81.3 _{0.3}	93.2 _{0.3}	81.3 _{0.3}	93.2 _{0.3}
DIFFPRUN-DEBIAS	84.5 _{0.1}	62.4 _{0.3}	84.2 _{0.4}	63.2 _{1.4}	81.6 _{0.4}	66.8 _{3.7}	81.3 _{0.2}	91.8 _{0.2}
MODDIFFY-POST	84.5 _{0.1}	61.6 _{0.7}	84.3 _{0.1}	64.5 _{0.4}	81.3 _{0.2}	66.0 _{1.2}	81.2 _{0.2}	91.1 _{0.9}
MODDIFFY-PAR	84.2 _{0.2}	53.7 _{1.5} ♣	84.5 _{0.2}	58.8 _{1.2} ♣	81.2 _{0.6}	75.4 _{3.7}	81.3 _{0.7}	85.5 _{0.4}

Table 1: Results of the BIOS and FCDL18 datasets on **BERT-Base** with adversarial bias removal and mutual information (MI) reduction. Task performance is measured with accuracy, and bias mitigation with balanced accuracy of the probes. The protected attribute is gender for BIOS, and race-dialect for FCDL18. The results with the best bias mitigation performance (lowest values) among the models that use *diff* subnetworks (lower part of the table) are shown in **bold**, and among all models with the ♣ symbol. Subscript values indicate standard deviation.

where the debiasing modules are learned after training the model. We use FINETUNE as the base model for MODDIFFY-POST, whose parameters are kept frozen during post-hoc training. **MODDIFFY-PAR**: our introduced parallel approach where the debiasing modules are learned together with the base model finetuned on the task. Additionally, to provide a comprehensive view on bias mitigation methods, we evaluate the datasets on the INLP (Ravfogel et al., 2020) approach using the implementation and suggested hyperparameter setting. The evaluation results of the INLP method are separately reported in Table 8 in Appendix B. As the PLM encoder for all models, we use two versions of BERT (Devlin et al., 2019) with different sizes, namely BERT-Mini (Turc et al., 2019) and BERT-Base. This provides us a more comprehensive picture regarding the effect of encoder size and number of involved parameters on the bias mitigation methods.

All debiasing models are separately trained according to the adversarial bias mitigation and mutual information reduction methods. We particularly opt for a non-linear adversarial head with two fully connected layers and the tanh activation. Furthermore, to improve the capacity of adversarial learning, we initialize 5 instances of h_{ρ_i} and calculate the average of the loss for the backward pass. In MI reduction, in the case of protected attributes with more than two classes, we turn the multi-classes setting to multiple one-versus-rest splits. For the models with *diff* subnetworks, we conduct preliminary experiments to find proper thresholds for magnitude pruning that improves

sparsity as much as possible without sacrificing performance. For BERT-Base and BERT-Mini, we set the minimum sparsity threshold to 99% and 95% (maximum size of 1% and 5%), respectively. We note that these ratios are the lower bounds, and the L_0 regularization may by itself reach a higher sparsity. The complete details of our hyperparameters setting and training procedure are explained in Appendix A. Our code and trained resources are available in <https://github.com/CPJKU/ModularizedDebiasing>.

Evaluation Metrics We evaluate the performance of the classifiers on the core task using the accuracy metric. We evaluate bias mitigation based on the concept of *fairness through blindness*, namely by examining whether models are agnostic about the protected attributes. Concretely following previous works (Elazar and Goldberg, 2018; Barrett et al., 2019), we report the leakage of a protected attribute in terms of the performance of a strong probe (or attacker) network. To train the probe, we freeze the model’s parameters, and train a new classification head (two-layer feed-forward layer with a tanh activation) to predict the protected attribute from the z encoding vector. For each evaluation, we train an ensemble of 5 probes for 40 epochs with early stopping if validation loss does not increase over 5 epochs, and report the results of the best performing probe. We report the performance of the probe in terms of balanced/macro accuracy (average of per-class accuracy scores). Balanced accuracy has the benefit of better reflecting the performance of the methods when considering minority groups, particularly given the unbalanced

Model	Adversarial Bias Mitigation			MI Reduction		
	Task↑	Probe _{G↓}	Probe _{A↓}	Task↑	Probe _{G↓}	Probe _{A↓}
FINETUNE	93.5 _{0.2}	70.4 _{2.2}	53.6 _{3.2}	93.5 _{0.2}	70.4 _{2.2}	53.6 _{3.2}
ADAPTER	87.3 _{0.1}	67.8 _{0.6}	37.1 _{1.7}	87.3 _{0.1}	67.8 _{0.6}	37.1 _{1.7}
FINETUNE-DEBIAS _G	93.7 _{0.0}	52.3 _{0.5}	34.8 _{1.1}	93.7 _{0.1}	62.2 _{0.1}	41.9 _{1.4}
FINETUNE-DEBIAS _A	92.9 _{0.2}	52.3 _{0.3}	31.4 _{2.6}	93.7 _{0.1}	61.5 _{0.7}	41.2 _{2.1}
FINETUNE-DEBIAS _{G&A}	93.0 _{0.2}	51.7 _{1.4}	33.9 _{1.8}	93.5 _{0.1}	61.4 _{0.7}	40.8 _{1.0}
ADAPTER-DEBIAS _G	86.9 _{0.1}	54.9 _{0.4}	29.2 _{1.7}	87.0 _{0.1}	67.4 _{0.3}	37.0 _{2.2}
ADAPTER-DEBIAS _A	86.1 _{0.2}	58.5 _{0.2}	25.6 _{2.0} ♣	87.1 _{0.3}	67.1 _{0.5}	36.6 _{1.2}
ADAPTER-DEBIAS _{G&A}	92.2 _{0.2}	51.6 _{2.4} ♣	27.5 _{1.9}	92.8 _{0.1}	64.2 _{0.4}	34.9 _{0.2}
DIFFPRUN	93.0 _{0.1}	76.1 _{0.4}	62.4 _{0.5}	93.0 _{0.1}	76.1 _{0.4}	62.4 _{0.5}
DIFFPRUN-DEBIAS _G	93.4 _{0.1}	56.3 _{0.8}	48.5 _{1.2}	93.3 _{0.0}	73.0 _{0.7}	58.2 _{0.5}
DIFFPRUN-DEBIAS _A	92.9 _{0.1}	64.5 _{0.8}	34.1 _{1.1}	93.4 _{0.0}	73.1 _{0.8}	57.1 _{0.6}
DIFFPRUN-DEBIAS _{G&A}	92.9 _{0.2}	54.7 _{2.6}	29.7 _{2.0}	93.6 _{0.2}	73.3 _{0.3}	57.2 _{0.9}
MODDIFFY-POST _G	93.7 _{0.1}	57.0 _{1.0}	45.4 _{1.3}	93.6 _{0.1}	66.2 _{0.2}	46.2 _{1.4}
MODDIFFY-POST _A	93.7 _{0.1}	63.7 _{0.9}	31.7 _{2.8}	93.6 _{0.0}	66.6 _{0.2}	46.5 _{2.8}
MODDIFFY-POST _G + <i>ditto</i> _A	92.3 _{0.6}	57.7 _{1.2}	32.0 _{2.8}	93.4 _{0.1}	66.4 _{0.8}	46.6 _{1.8}
MODDIFFY-POST _{G&A}	93.6 _{0.1}	52.6 _{2.2}	30.9 _{0.5}	93.6 _{0.1}	66.9 _{0.4}	47.4 _{0.1}
MODDIFFY-PAR _G	93.5 _{0.2}	53.0 _{1.6}	32.2 _{1.1}	93.6 _{0.0}	59.3 _{0.4}	35.7 _{0.4}
MODDIFFY-PAR _A	93.5 _{0.2}	53.8 _{1.9}	30.1 _{1.0}	93.7 _{0.2}	55.0 _{0.2}	29.9 _{0.9}
MODDIFFY-PAR _G + <i>ditto</i> _A	93.5 _{0.2}	52.8 _{1.7}	30.2 _{0.8}	93.6 _{0.2}	56.1 _{1.1}	30.1 _{1.0}
MODDIFFY-PAR _{G&A}	93.8 _{0.2}	52.3 _{1.4}	28.3 _{1.6}	93.6 _{0.2}	52.7 _{1.3} ♣	29.4 _{0.6} ♣

Table 2: Results of the PAN16 dataset on **BERT-Base**. The subscripts G and A refer to the protected attributes gender and age, respectively. The sign G&A denotes that the bias mitigation loss is the sum of the debiasing loss terms of both gender and age. The MODDIFFY models in the form of “+*ditto*” refer to the case where first two debiasing subnetworks with the same core model are trained separately, and then they are added to the base model at inference time. The results with the best bias mitigation performance (lowest values) among the models that use *diff* subnetworks (lower part of the table) are shown in **bold**, and among all models with the ♣ symbol. Subscript values indicate standard deviation.

distributions over protected labels in the datasets. To account for possible variabilities, we repeat every experiment five times and report the mean and standard deviation.

5 Results and Analysis

Single-attribute Evaluation Table 1 reports the evaluation results of the BIOS and FCDL18 datasets on BERT-Base using adversarial bias removal and mutual information (MI) reduction. The results of the same experiments on BERT-Mini are shown in Table 6 in Appendix B.

Starting from task accuracy, the models using subnetworks (variations of DIFFPRUN and MODDIFFY shown at the lower part of the table) consistently perform the same as the fully finetuned models on both datasets and debiasing methods. We observe a slight decrease in performance for the adapter-based models on FCDL18.

Looking at the leakage probing performance of bias attributes, MODDIFFY models show better

performance (lower values) among the subnetwork-based models on both datasets, and overall on BIOS.² In particular, MODDIFFY models outperform the directly comparable baselines FINETUNE-DEBIAS and DIFFPRUN-DEBIAS on all configurations, indicating the benefits of learning separate debiasing modules on bias mitigation performance. This indeed comes with the core advantages of MODDIFFY models in proving modularized and on-demand bias mitigation.

The results also show that MODDIFFY-POST, while (as expected) slightly weaker than MODDIFFY-PAR, provides competitive bias mitigation performance (i.e., on par with DIFFPRUN-DEBIAS). Finally, comparing between debiasing optimizations, MI reduction shows consistently worse bias

²Particularly on FCDL18, we observe high variations, which requires us to more cautiously interpret results. We assume that this is due to the small size of this dataset especially in the learning regimes with many parameters on BERT-Base, as this effect is less pronounced on BERT-Mini (Table 6 in Appendix B).

mitigation performance in comparison with adversarial training, particularly on FCDL18 where the protected attribute has more than two labels.

Two-attribute Evaluation The results on PAN16 using BERT-Base are reported in Table 2, and the same experiments on BERT-Mini in Table 7 in Appendix B. In this experiment, every debiasing model is trained based on either gender, age, or simultaneously on both gender and age, shown with the subscripts G, A, G&A, respectively. An additional evaluation is indicated with MODDIFFY-*_G + *ditto*_A, which refers to adding the two separately trained gender and age debiasing subnetworks to the core model at inference time. In fact, for the experiments of each MODDIFFY model indicated with “+*ditto*”, G, and A, we train only one model with gender and age subnetworks, and then add the respective subnetwork(s) to the core model.

Looking at task performance results, similar to the previous datasets the models perform on par with fully finetuning, except the adapter-based models which in this case significantly underperform in both optimization methods. Regarding bias mitigation performance, we observe similar patterns to the ones discussed on the other datasets: MODDIFFY models particularly MODDIFFY-PAR show the least attribute leakage among the subnetwork-based models consistently, and also over all models with MI reduction. This reaffirms the benefits of modularizing debiasing of the attributes separately from the task. The results also indicate the existence of a correlation between the gender and age attributes in this dataset, such that debiasing each attribute also results in a decrease in leakage of the other attribute. The overall best results are achieved on the models that simultaneously optimize on both attributes (G&A).

Finally, let us have a closer look at the results of applying the two independently trained subnetworks. On both training regimes, we observe on par debiasing performance between MODDIFFY-*_G + *ditto*_A and the corresponding results with one subnetwork, namely the ones of gender and age debiasing in MODDIFFY-*_G and MODDIFFY-*_A, respectively. These results indicate the viability of our approach to effectively merge subnetworks at inference time.

Subnetworks analysis We further investigate the achieved sparsity rate of the subnetworks, keeping

	PAN16 Gender	PAN16 Age	BIOS Gender	FCDL18 Dialect
Overall	0.01%	1.00%	0.27%	0.18%
Layer 12	0.04%	7.00%	0.26%	0.92%
Layer 11	0.02%	3.59%	0.30%	0.52%
Layer 10	0.01%	1.76%	0.33%	0.30%
Layer 9	0.00%	0.49%	0.20%	0.18%
Layer 8	0.00%	0.38%	0.22%	0.11%
Layer 7	0.01%	0.21%	0.30%	0.14%
Layer 6	0.00%	0.15%	0.28%	0.13%
Layer 5	0.01%	0.03%	0.28%	0.07%
Layer 4	0.01%	0.25%	0.43%	0.08%
Layer 3	0.01%	0.25%	0.56%	0.08%
Layer 2	0.01%	0.38%	0.41%	0.07%
Layer 1	0.01%	0.21%	0.44%	0.06%
Embeddings	0.00%	0.04%	0.05%	0.03%

Table 3: The percentage of non-masked parameters in the debiasing subnetworks of MODDIFFY-PAR.

in mind that the maximum capacity of the debiasing subnetworks on BERT-Base is set to 1%. The size of a subnetwork indicates the amount of information or in fact modifications needed to be applied, in order to debias a protected attribute. Table 3 reports the percentage of the number of non-masked parameters in every layer, and also overall, in the subnetworks of the MODDIFFY-PAR models regarding the protected attributes. The results show interesting patterns in respect to various protected attributes: the gender attribute on both PAN16 and BIOS dataset require much smaller subnetworks, such that the subnetwork on PAN16 is only 0.01% of the size of the core model. The age attribute appears to be a more complex topic in the underlying PLM, as it fully uses the 1% maximum capacity. Looking across the layers, the results show that debiasing the gender attribute is mostly handled in the lower transformer layers (particularly on BIOS), while debiasing age and dialect attributes mostly happens at the higher layers.

In Appendix C, we further discuss this topic for all models on the level of individual weight matrices. Moreover, in Appendix D we investigate to what extent the subnetworks of a model across several runs affect on the same set of parameters.

6 Conclusion

We propose MODDIFFY, a novel bias mitigation approach which enables integration of an arbitrary subset of the debiasing modules at inference time. Our method encapsulates the functionality of bias mitigation in respect to a protected attribute into a separate magnitude-difference subnetwork, which can then be applied to the core model on-demand. Our experiments on three classification tasks show

that MODDIFFY improves bias mitigation achieved by separating the debiasing from the task network, and effectively mitigates the bias of two (and possibly more) attributes when their respective subnetworks are simultaneously utilized.

7 Limitations

An important limitation of our work concerns the definition of the protected attributes in the datasets used for evaluation. In particular, gender in BIOS and PAN16 is limited to the binary female/male, lacking an inclusive and nuanced definition of gender. Similarly in FDCL18, we consider only two dialects of *African American* and *White American*, while clearly this definition is limited and non-inclusive. Furthermore as in previous work (Sap et al., 2019; Ravfogel et al., 2020; Zhang et al., 2021), the labels of this protected attribute are assigned through a probabilistic model, and hence the dataset might not represent the nuances and traits of the real-world.

The second limitation regards reaching strong conclusions on the generalizability of the multi-attribute setting for MODDIFFY over any possible number of protected attributes or subset of them. Our multi-attribute experiments are conducted on one dataset with two attributes of gender and age, particularly due to the lack of available suitable datasets. Hence, Further studies (as well as more suitable datasets) are required for achieving a more comprehensive picture on the topic.

Finally, we should also highlight two general limitations, shared with the other related studies in the area of model bias mitigation. First, we should consider that the aim of representation disentanglement optimizations is to reduce the existing *correlations* in the model with the protected attributes based on the *observed data*. These data-oriented approaches might lack effective generalization, particularly when the model is evaluated in other domains or out-of-distribution data. Second, our bias mitigation evaluation is grounded in the notion of *fairness through blindness*, and the debiasing optimization methods are designed to support this form of fairness. The effects of our method on other possible definitions of fairness are therefore left for future work.

8 Acknowledgment

This work received financial support by the Austrian Science Fund (FWF): P33526 and DFH-23;

and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grants LIT-2020-9-SEE-113 and LIT-2021-YOU-215.

References

- Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6331–6336.
- Elad Ben-Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *ArXiv*, abs/2106.10199.
- Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 405–414.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.
- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to home-maker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *Proceedings of NeurIPS*.
- Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541, Online. Association for Computational Linguistics.
- Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021. A novel estimator of mutual information for learning to disentangle textual representations. 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 6539–6550.
- 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*. page 120–128, New York, NY, USA. Association for Computing Machinery.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding*. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186. Association for Computational Linguistics.
- Yanai Elazar and Yoav Goldberg. 2018. *Adversarial removal of demographic attributes from text data*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21. Association for Computational Linguistics.
- Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.
- Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of International Conference on Learning Representations, ICLR*.
- Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. 2022. *Unlearning protected user attributes in recommendations with adversarial training*. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’22*, page 2142–2147, New York, NY, USA. Association for Computing Machinery.
- Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. Proceedings of Machine Learning Research.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. In *Journal of Machine Learning Research*, volume 17, pages 1–35.
- Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.
- Demi Guo, Alexander Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4896.
- Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.
- Wenjuan Han, Bo Pang, and Ying Nian Wu. 2021a. Robust transfer learning with pretrained language models through adapters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 854–861.
- Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021b. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, volume 97, pages 2790–2799. Proceedings of Machine Learning Research.
- Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*.
- Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Łukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. 2021. Sparse is enough in scaling transformers. *Advances in Neural Information Processing Systems*, 34.
- Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.
- Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2023. Grep-biasir: A dataset for investigating gender representation-bias in information retrieval results. In *Proceeding of the ACM SIGIR Conference On Human Information Interaction And Retrieval (CHIIR)*.
- François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. 2021. Block pruning for faster transformers. In *Proceedings of the 2021 Conference*

- on Empirical Methods in Natural Language Processing*, pages 10619–10629.
- Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. **Sustainable modular debiasing of language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Christos Louizos, Max Welling, and Diederik P. Kingma. 2018. Learning sparse neural networks through ℓ_0 regularization. In *International Conference on Learning Representations*.
- David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *Proceedings of the International Conference on Machine Learning*, pages 3384–3393. PMLR.
- Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. **Compacter: Efficient low-rank hypercomplex adapter layers**. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 1022–1035.
- Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in NLU. In *Proceeding of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: Cross-genre evaluations. In *CLEF*.
- Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.
- Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. **Pan16 author profiling**. Zenodo.
- 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.
- Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of BERT rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–316.
- Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2065–2068.
- Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. **The risk of racial bias in hate speech detection**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- 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.
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3398–3403.
- 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.
- Asa Cooper Stickland and Iain Murray. 2019. **BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning**. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.
- Yi-Lin Sung, Varun Nair, and Colin Raffel. 2021. Training neural networks with fixed sparse masks. *ArXiv*, abs/2111.09839.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.
- Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

- Liwen Wang, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu. 2021. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3740–3750.
- Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. 2020. Supermasks in superposition. *Advances in Neural Information Processing Systems*, 33.
- Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.
- Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 585–596.
- Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’22, page 2532–2538, New York, NY, USA. Association for Computing Machinery.
- Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. 2021. Disentangling representations of text by masking transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 778–791, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiayu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 629–634.
- Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Empirical Methods in Natural Language Processing*, pages 2226–2241. Association for Computational Linguistics.
- Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. 2019. Deconstructing lottery tickets: Zeros, signs, and the supermask. *Advances in Neural Information Processing Systems*, 32:3597–3607.

A Experiment Settings – Additional Details

In FDCL18 dataset, we use the TwitterAAE model (Blodgett et al., 2016) to assign racial dialect classes. The TwiiterAAE model predicts four racial classes, *African American*, *White American*, *Hispanic*, and *Others*. We labeled a tweet as *African American* or *White American* if the prediction score was greater than 0.5. For PAN16 dataset, following (Sap et al., 2019) we balanced the task labels and sampled 200K data. The age groups of this dataset are 18-24, 25-34, 35-49, 50-64, and 65+.

We randomly split the dataset into train, validation, and test set with the proportions 63:12:15 for BIOS, 63:12:15 for FDCL18, and 80:5:15 on PAN16. We use the validation set for hyperparameter tuning, and the best result on the validation set is evaluated on test set for the final results. The validation and test sets in all datasets follow the same distribution as the whole dataset. To address the unbalancedness of the dataset and the potential problems in adversarial learning, we apply upsampling only on the *training sets* of BIOS and FDCL18 datasets, to balance the protected attribute labels within each task label. For instance, genders are balanced in the dentist class by repeating the data items of the minority subgroup.

Adversarial heads consist of five classifiers with different initialization. The loss for the five classifiers is averaged and accuracy is measured via majority vote. All baseline models are trained for 20 epochs. All DIFFPRUN and MODDIFFY model variants are trained for 30 epochs as they need to account for the two phases of *diff* pruning, where a model requires more training to recover its performance after the magnitude pruning step. We fix the learning rate of BERT weights to $2e-5$ and the learning rate for the classifier heads to $1e-4$. We set the batch size to 64 for all experiments. We keep other *diff*-specific hyperparameters the same as suggested by Guo et al. (2021). Adapter baselines follow Pfeiffer et al. (2021) with reduction factor of two. Rest of the hyperparameters are same as the subnetwork-based models. Table 4 reports the hyperparameters of our experiments.

B Additional Results

The results of all models using BERT-Mini are shown in Table 6 for BIOS and FCDL18 datasets, and in Table 7 for PAN16. Table 8 reports the evaluation results of the INLP method.

training	
batch_size	64
structured_diff_pruning	True
alpha_init	5
concrete_samples	1
concrete_lower	-1.5
concrete_upper	1.5
num_epochs	20
num_epochs_finetune	15
num_epochs_fixmask	15
learning_rate	2e-05
learning_rate_task_head	0.0001
learning_rate_adv_head	0.0001
learning_rate_alpha	0.1
task_dropout	0.3
task_n_hidden	0
adv_dropout	0.3
adv_n_hidden	1
adv_count	5
adv_lambda	1.0
sparsity_pen	1.25e-07
max_grad_norm	1.0
adv attack	
batch_size	64
num_epochs	40
learning_rate	0.0001
adv_n_hidden	1
adv_count	5
adv_dropout	0.3

Table 4: Hyperparameters used for training

Parameter Name	Size
word_embeddings	117,204,480
intermediate.dense	11,811,840
output.dense	11,800,320
attention.self.query	2,952,960
attention.self.key	2,952,960
attention.self.value	2,952,960
attention.output.dense	2,952,960
position_embeddings	1,966,080
output.adapter	590,976
others	7680

Table 5: BERT-Base number of parameters

C Sparsity rates of subnetworks

We visualize the percentage of non-masked parameters of the subnetworks in BERT-Base for each parameter matrix in Figures 2, 3 and 4 for MODDIFFY-PAR, MODDIFFY-POST, and DIFFPRUN-DEBIAS, respectively. In addition to the discussion in Section 5, we observe in these detailed figures that the LayerNorm module of the last Transformer block generally has a high density. We assume

Model	BIOS				FCDL18			
	Adversarial		MI Reduction		Adversarial		MI Reduction	
	Task↑	Probe↓	Task↑	Probe↓	Task↑	Probe↓	Task↑	Probe↓
FINETUNE	82.9 _{0.1}	65.5 _{0.4}	82.9 _{0.1}	65.5 _{0.4}	82.1 _{0.2}	90.3 _{0.9}	82.1 _{0.2}	90.3 _{0.9}
ADAPTER	81.6 _{0.2}	65.7 _{0.2}	81.6 _{0.2}	65.7 _{0.2}	81.9 _{0.0}	79.2 _{0.5}	81.9 _{0.0}	79.2 _{0.5}
FINETUNE-DEBIAS	81.6 _{0.2}	56.4 _{1.7}	81.6 _{0.1}	59.9 _{0.9}	80.0 _{0.5}	73.7 _{2.7}	79.7 _{0.5}	87.3 _{2.6}
ADAPTER-DEBIAS	81.5 _{0.1}	63.4 _{0.1}	81.7 _{0.1}	65.4 _{0.1}	81.1 _{0.2}	64.6 _{1.3} ♣	81.8 _{0.1}	78.5 _{0.5} ♣
DIFFPRUN	83.5 _{0.1}	65.8 _{0.3}	83.5 _{0.1}	65.8 _{0.3}	82.8 _{0.1}	92.6 _{0.9}	82.8 _{0.1}	92.6 _{0.9}
DIFFPRUN-DEBIAS	83.3 _{0.1}	59.1 _{0.7}	82.3 _{0.1}	65.5 _{0.9}	82.3 _{0.4}	65.8 _{2.8}	82.5 _{0.3}	91.9 _{0.9}
MODDIFFY-POST	83.1 _{0.0}	57.3 _{0.7}	83.1 _{0.0}	63.4 _{1.4}	81.6 _{0.4}	69.7 _{1.9} ♣	82.3 _{0.1}	89.3 _{0.4}
MODDIFFY-PAR	81.5 _{0.2}	55.7 _{0.6} ♣	81.4 _{0.2}	58.8 _{0.8} ♣	80.2 _{0.3}	73.8 _{6.7}	79.9 _{8.6}	85.4 _{1.6}

Table 6: Results of the BIOS and FCDL18 datasets on **BERT-Mini** with adversarial bias removal and mutual information (MI) reduction. Task performance is measured with accuracy, and bias mitigation with balanced accuracy of the probes. The protected attribute is gender for BIOS, and race-dialect for FCDL18. The results with the best bias mitigation performance (lowest values) among the models that use *diff* subnetworks (lower part of the table) are shown in **bold**, and among all models with the ♣ symbol.

Model	Adversarial Bias Mitigation			MI Reduction		
	Task↑	Probe _G ↓	Probe _A ↓	Task↑	Probe _G ↓	Probe _A ↓
FINETUNE	91.5 _{0.2}	64.8 _{0.6}	46.8 _{0.3}	91.5 _{0.2}	64.8 _{0.3}	88.4 _{1.1}
ADAPTER	78.4 _{0.2}	65.8 _{0.2}	35.3 _{0.8}	78.4 _{0.2}	65.8 _{0.2}	35.3 _{0.8}
FINETUNE-DEBIAS _G	91.7 _{0.2}	54.6 _{0.6}	42.2 _{0.6}	91.6 _{0.2}	61.3 _{0.6}	42.4 _{0.5}
FINETUNE-DEBIAS _A	91.1 _{0.2}	61.9 _{0.6}	39.1 _{0.8}	91.4 _{0.6}	61.9 _{0.1}	43.2 _{0.0}
FINETUNE-DEBIAS _{G&A}	91.2 _{0.1}	57.0 _{1.0}	38.5 _{0.6}	91.9 _{0.0}	62.5 _{0.1}	42.1 _{0.9}
ADAPTER-DEBIAS _G	78.1 _{0.1}	59.6 _{0.4}	32.0 _{1.3}	78.5 _{0.2}	65.9 _{0.2}	34.8 _{0.0}
ADAPTER-DEBIAS _A	77.3 _{0.1}	60.5 _{0.9}	27.3 _{0.9}	78.3 _{0.1}	65.5 _{0.2}	34.1 _{0.3} ♣
ADAPTER-DEBIAS _{G&A}	80.9 _{0.6}	55.6 _{0.4}	25.5 _{1.2} ♣	82.0 _{0.1}	64.2 _{0.4}	34.9 _{0.2}
DIFFPRUN	90.0 _{0.1}	67.2 _{0.3}	49.4 _{0.8}	90.0 _{0.1}	67.2 _{0.3}	49.4 _{0.8}
DIFFPRUN-DEBIAS _G	90.1 _{0.1}	54.1 _{1.1} ♣	44.1 _{0.7}	78.5 _{0.2}	65.9 _{0.2}	34.8 _{0.0}
DIFFPRUN-DEBIAS _A	89.2 _{0.3}	64.8 _{0.6}	39.4 _{1.9}	78.3 _{0.1}	65.5 _{0.2}	34.1 _{0.3} ♣
DIFFPRUN-DEBIAS _{G&A}	89.1 _{0.1}	57.9 _{1.7}	35.8 _{2.2}	86.3 _{0.0}	68.8 _{0.1}	52.3 _{0.9}
MODDIFFY-POST _G	91.7 _{0.2}	55.7 _{1.2}	42.7 _{0.6}	91.5 _{0.0}	62.6 _{0.0}	43.7 _{1.3}
MODDIFFY-POST _A	91.4 _{0.2}	62.3 _{0.5}	34.8 _{0.8}	91.5 _{0.0}	62.8 _{0.3}	44.0 _{0.0}
MODDIFFY-POST _G + <i>ditto</i> _A	91.0 _{0.4}	59.3 _{0.6}	37.3 _{1.2}	!!! 91.5 _{0.5}	62.8 _{0.4}	43.9 _{0.8}
MODDIFFY-POST _{G&A}	91.5 _{0.2}	56.7 _{1.3}	35.1 _{2.1}	91.5 _{0.0}	63.0 _{0.3}	43.9 _{0.4}
MODDIFFY-PAR _G	91.6 _{0.2}	55.9 _{0.8}	41.7 _{0.7}	91.6 _{0.1}	60.9 _{0.4}	40.3 _{0.8}
MODDIFFY-PAR _A	91.3 _{0.2}	61.3 _{0.5}	37.6 _{1.2}	91.5 _{0.2}	60.8 _{0.6}	41.6 _{0.7}
MODDIFFY-PAR _G + <i>ditto</i> _A	91.4 _{0.4}	60.7 _{1.0}	39.6 _{1.6}	91.7 _{0.1}	60.2 _{0.5} ♣	39.7 _{0.1}
MODDIFFY-PAR _{G&A}	91.3 _{0.3}	55.5 _{1.1}	33.7 _{1.6}	91.5 _{0.3}	60.7 _{1.3}	41.3 _{1.1}

Table 7: Results of the PAN16 dataset on **BERT-Mini**. The subscripts G and A refer to the protected attributes gender and age, respectively. The sign G&A denotes that the bias mitigation loss of the model consists of the debiasing loss terms of both gender and age. The MODDIFFY models in the form of “+*ditto*” refer to the case, where first two debiasing subnetworks with the same core model are trained separately, and then they are added to the base model at inference time. The results with the best bias mitigation performance (lowest values) among the models that use *diff* subnetworks (lower part of the table) are shown in **bold**, and among all models with the ♣ symbol.

Model	BIOS (gender)		FDCL18 (race)	
	Task↑	Probe↓	Task↑	Probe↓
BERT-Mini	71.1 _{0.2}	59.8 _{0.9}	73.6 _{1.6}	57.9 _{0.4}
BERT-Base	66.2 _{0.4}	50.5 _{0.3}	71.6 _{2.2}	50.2 _{0.1}

(a) BIOS and FDCL18

Model	Task↑	Probe _G ↓	Probe _A ↓
BERT-Mini	<i>G</i>	69.3 _{0.1}	60.1 _{0.3}
	<i>A</i>	66.7 _{1.8}	60.6 _{0.1}
BERT-Base	<i>G</i>	69.4 _{0.1}	54.4 _{0.1}
	<i>A</i>	49.8 _{0.8}	54.6 _{0.6}

(b) PAN16

Table 8: Evaluation results of INLP.

that this is due to the additive nature of these *diff*-based methods, as changing the weight magnitude through adding a subnetwork requires rescaling the final output.

D Consistency in finding subnetworks

Figures 5, 6 and 7 show the percentage of common non-masked parameters in subnetworks on a particular weight matrix/layer across 5 runs for MODDIFFY-PAR, MODDIFFY-POST, and DIFF-PRUN-DEBIAS, respectively. We report the percentage of overlap between the subnetworks of two, three, four, and five runs, separated with the “/”. The results show that the equivalent subnetworks across various runs (with different initialization seeds) seem to be largely separated. This results are consistent with the observations on the lottery ticket hypothesis on large neural networks (Chen et al., 2020).

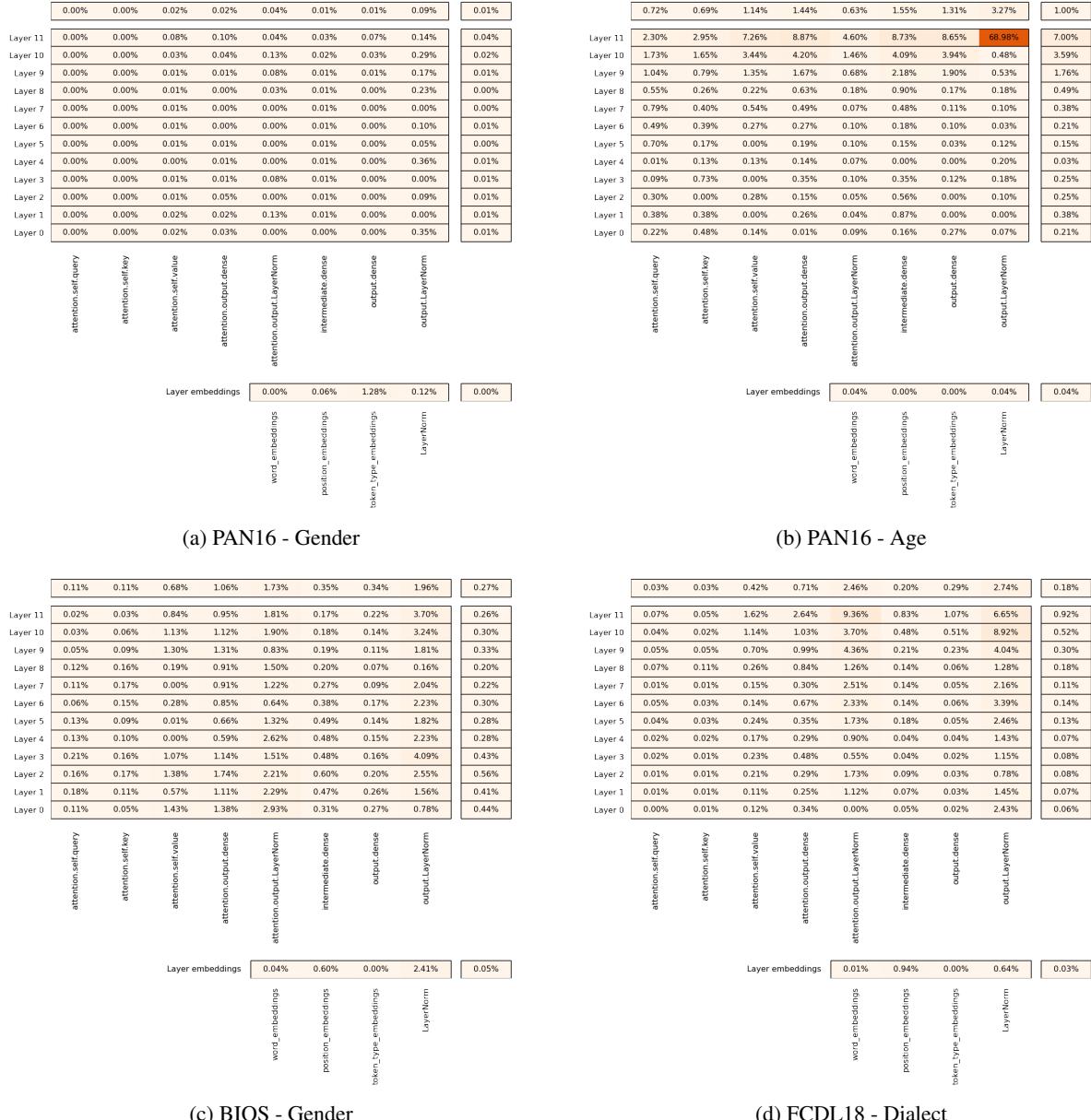


Figure 2: Sparsity rate of the subnetworks in MODDIFFY-PAR on BERT-Base. Each value shows the sparsity rate on the specific block/matrix.

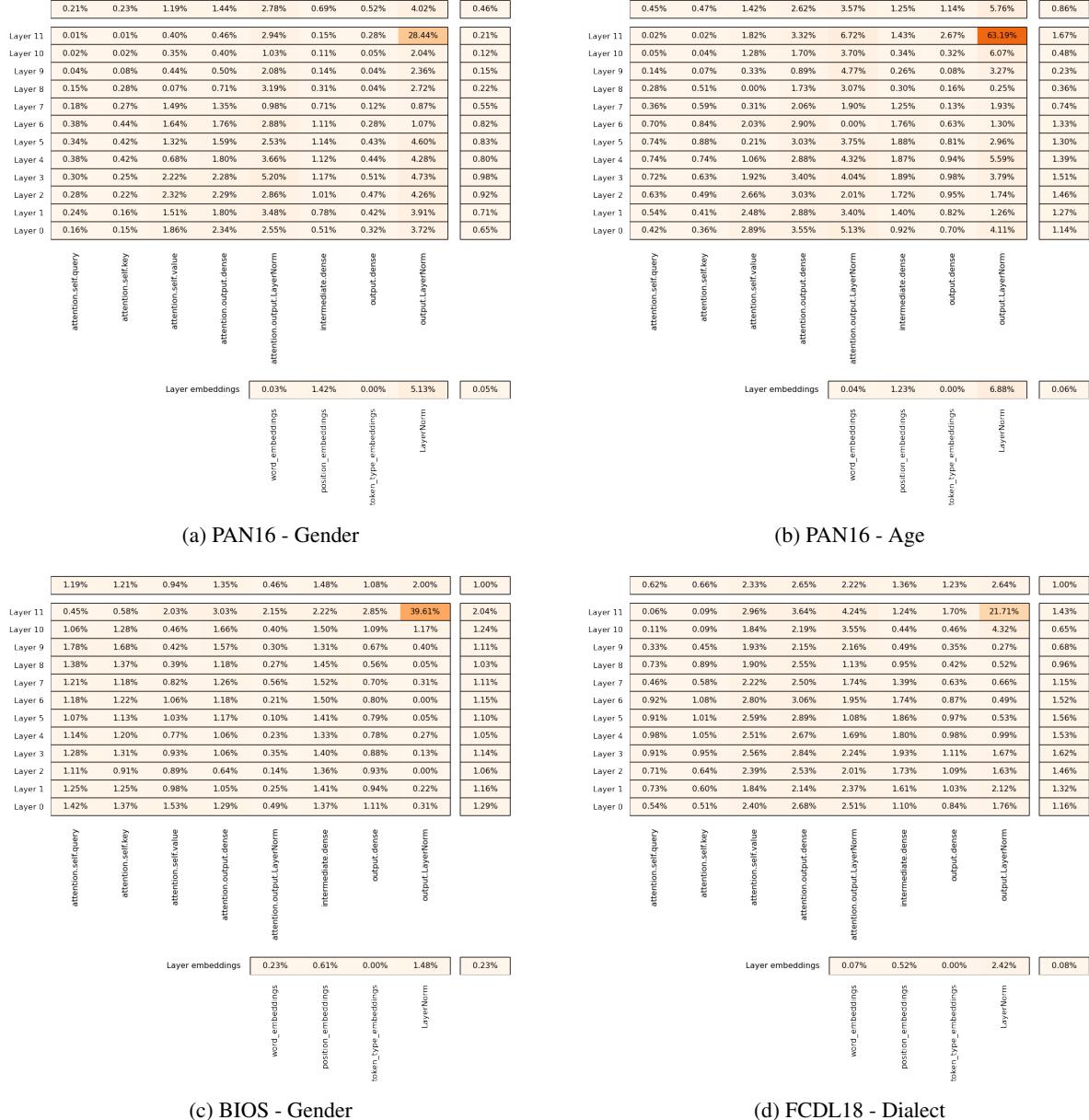


Figure 3: Sparsity rate of the subnetworks in MoDDIFFY-POST on BERT-Base. Each value shows the sparsity rate on the specific block/matrix.

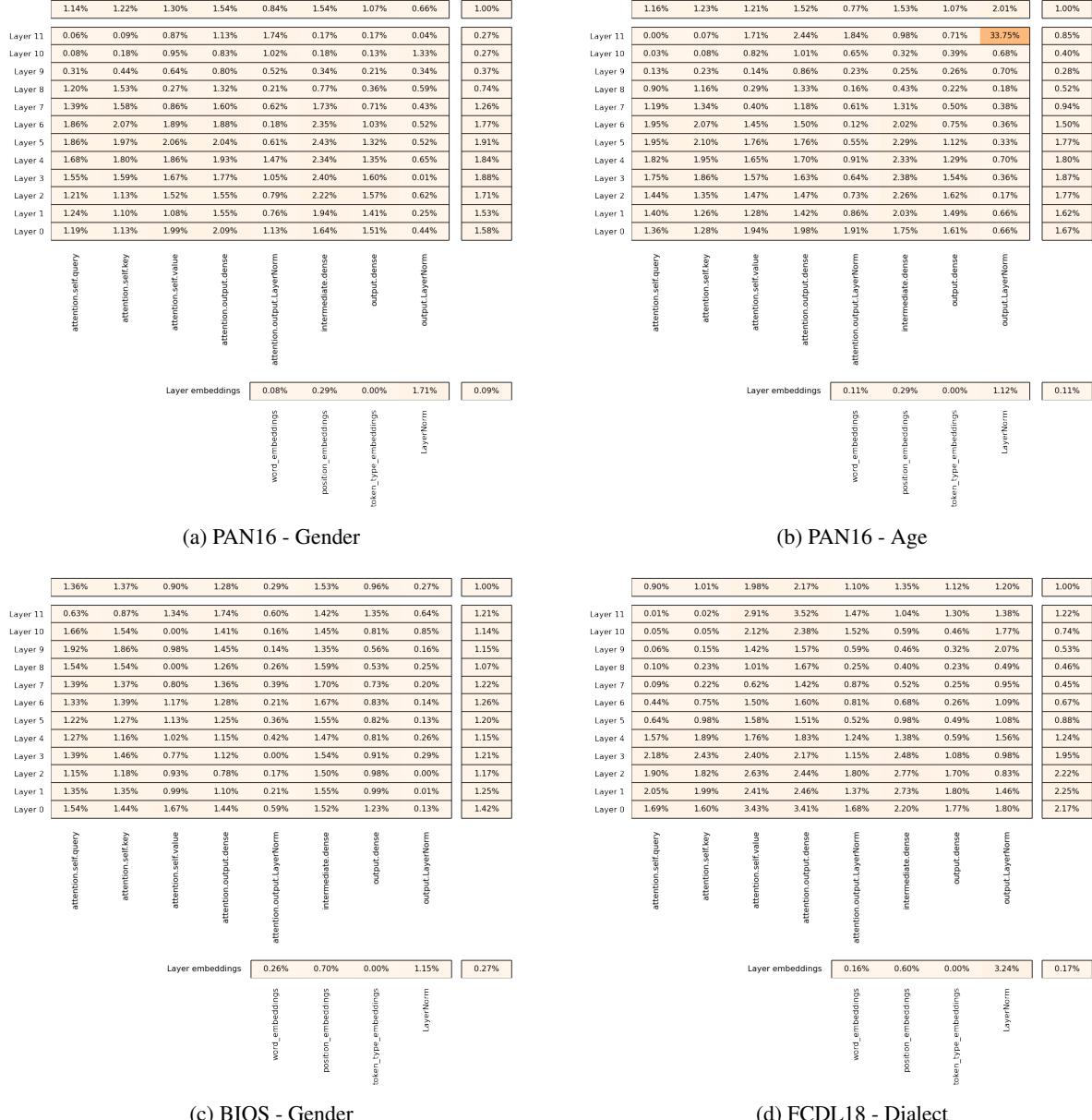


Figure 4: Sparsity rate of the subnetworks in DIFFPRUN-DEBIAS using BERT-Base. Each value shows the sparsity rate on the specific block/matrix.



Figure 5: The percentage of common non-masked parameters for MODDIFFY-PAR across 5 runs with different initialization seeds. For each block, the numbers indicate the percentage of the number of common parameters

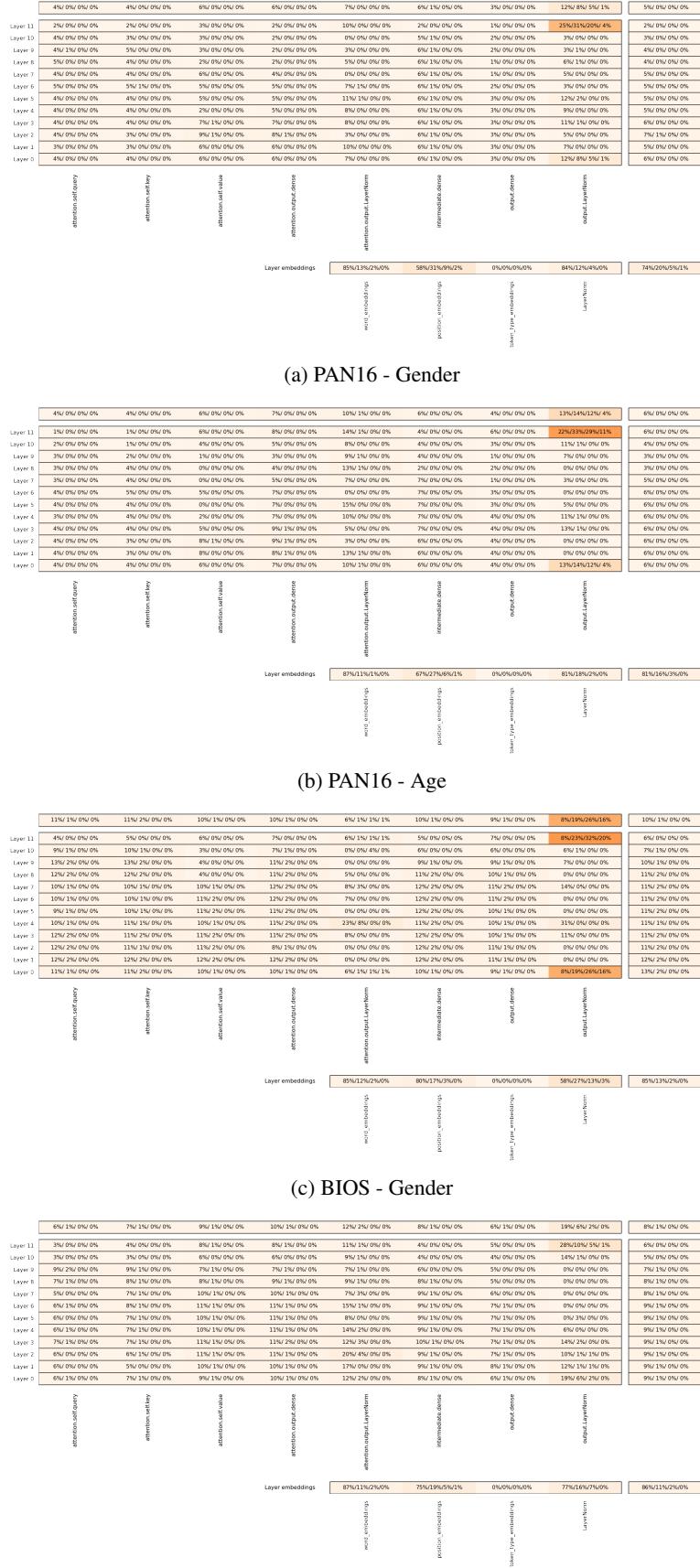


Figure 6: The percentage of common non-masked parameters for MoDDIFFY-POST across 5 runs with different initialization seeds. For each block, the numbers indicate the percentage of the number of common parameters across two, three, four, and five runs of a subnetwork, respectively.

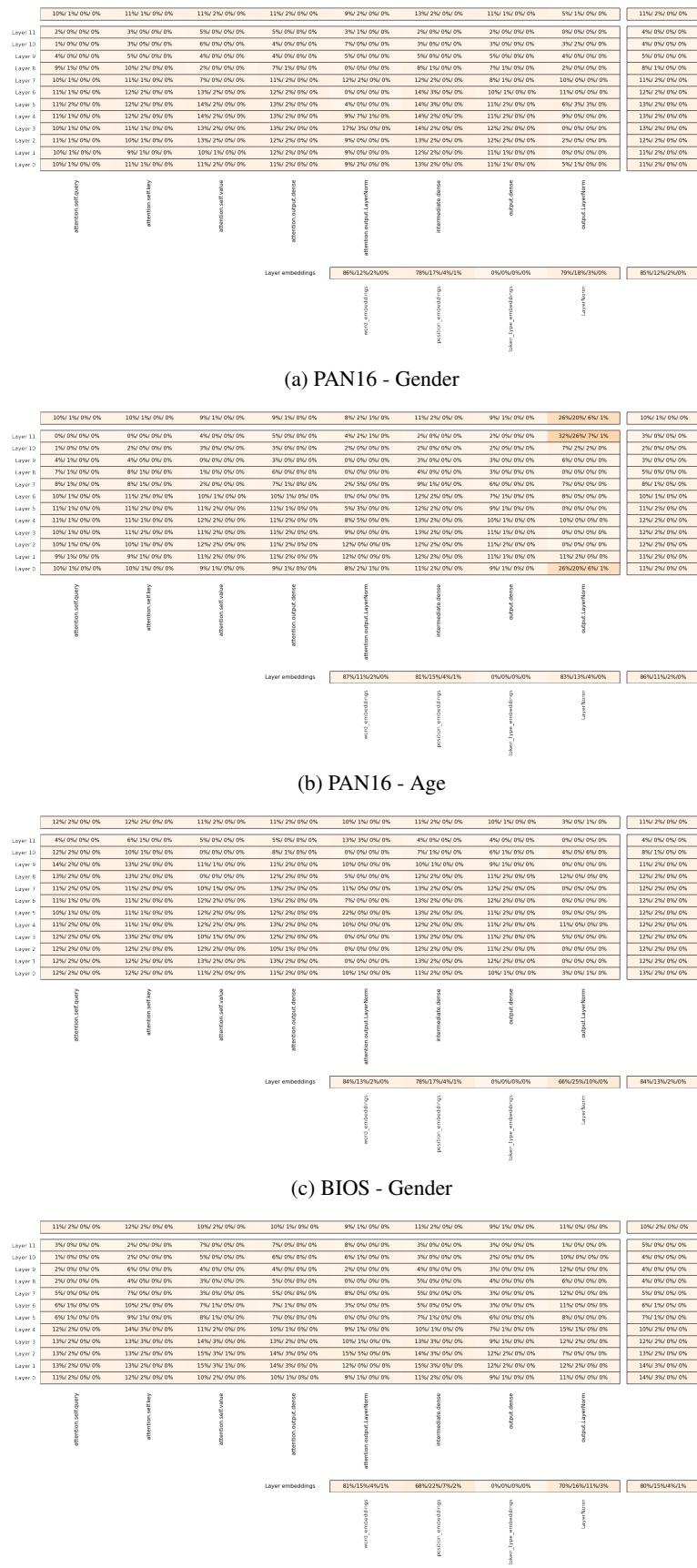


Figure 7: The percentage of common non-masked parameters for DIFFPRUN-DEBIAS across 5 runs with different initialization seeds. For each block, the numbers indicate the percentage of the number of common parameters

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Left blank.
- A2. Did you discuss any potential risks of your work?
Left blank.
- A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Section 4, 7 and Appendix A

- B1. Did you cite the creators of artifacts you used?
Section 4
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4, 7 and Appendix A
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 4 and Appendix A
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 and Appendix A
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4 and Appendix A

C Did you run computational experiments?

Section 4, 5, and Appendix A

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 4 and Appendix A

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Section 5, and Appendix B

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

No response.

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.