EFFICIENT LEARNING OF LESS BIASED MODELS WITH TRANSFER LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prediction bias in machine learning models, referring to undesirable model behaviors that discriminates inputs mentioning or produced by certain group, has drawn increasing attention from the research community given its societal impact. While a number of bias mitigation algorithms exist, it is often difficult and/or costly to apply them to a large number of downstream models due to the challenges on (sensitive) user data collection, expensive data annotation, and complications in algorithm implementation. In this paper, we present a new approach for creating less biased downstream models: transfer learning from a less biased upstream model. A model is trained with bias mitigation algorithms in the source domain and fine-tuned in the target domain without bias mitigation. By doing so, the framework allows to achieve less bias on downstream tasks in a more efficient, accessible manner. We conduct extensive experiments with the proposed framework under different levels of similarities between the source and target domain and the number of factors included for de-biasing. The results are positive, implying that less biased models can be obtained with our transfer learning framework.[1]

1 INTRODUCTION

Bias mitigation in machine learning models has drawn increasing attention in the natural language processing (NLP) community recently (Bolukbasi et al., 2016; Caliskan et al., 2017; Blodgett et al., 2020). Despite the strong results achieved from recent pretrained language models (PTLM) such as BERT (Devlin et al., 2019), the downstream classifiers still yield biased predictions for certain user groups (Zhao et al., 2019). Among many examples, Kurita et al. (2019) demonstrates that models trained using BERT embeddings for pronoun resolution are gender-biased. The models associate female entities with certain labels even when the label is associated with the same number female and male instances in the training set. Kennedy et al. (2020) shows that hate speech classifiers fine-tuned from BERT spuriously correlate certain group name mentions with "hate" label, resulting in false positive model predictions when group identifiers (*e.g.*, "*muslim*", "*black*") are present–*e.g.*, misclassifying non-hate speech text such as "*I am a Muslim*" as hate speech.

More recently, several bias-mitigation methods have been proposed (Park et al., 2018; Zhang et al., 2018; Beutel et al., 2017) to make a downstream classifier less biased in order to avoid discrimination (Mehrabi et al., 2019). However, these algorithms are designed specifically for training downstream classifiers: the bias mitigation algorithm applies *repeatedly* at the time of training each downstream classifier. It is costly in practice especially when there are often a number of downstream classifiers, as it is often difficult and/or costly to collect and annotate (sensitive) user data, and implementing mitigation algorithms. A potential solution is to look into the upstream models (*e.g.*, PTLMs), where a large number of downstream classifiers are derived. However, while there exist works on mitigating bias in PTLMs (Liang et al., 2020; Bhardwaj et al., 2020), they study bias in the representation space (*e.g.*, examining representational distances between gender and stereotypical words) or on downstream tasks while keeping representations are frozen; few study the capability of bias mitigation in fine-tuned downstream classifiers by manipulating upstream models.

In this paper, we propose the framework of upstream bias mitigation in transfer learning: an upstream model (source model) is trained with bias mitigation objectives in a source domain; then any downstream classifier can be built upon the encoder of the upstream model via fine-tuning without

[1] Link to code and data: `https://anonymous.4open.science/r/7e1ac8a0-d89a-4dca-8da8-30c03490fa42/`

additional bias mitigation process. The frameworks fits in the practice of fine-tuning from PTLMs, and allows to achieve less biases in downstream classifiers in a more light-weight, accessible manner. However, whether mitigation of bias is preserved in fine-tuning is not certain: downstream classifiers may ignore any prior knowledge about bias mitigation in the source model and pick up bias from data at fine-tuning.

We conduct a series of experiments on diverse settings and datasets to validate the capability of the framework: (1) in the most straightforward setting, the target domain consists of new examples from the same distribution as the source domain, which corresponds to fine-tuning a less biased model over emerging new examples; (2) in a more challenging setting, the source and the target domain have different data distributions (*e.g.*, different datasets); (3) furthermore, we perform multi-task learning to reduce different kinds of bias in the source domain, so that fine-tuned classifiers are less biased in term of multiple bias factors. We achieve overall positive empirical results, evidencing the capability of learning less biased models with the proposed upstream bias mitigation framework.

2 Related Works

Bias, in our setting, refers to undesirable model behaviors, such as spuriously correlated mentions of certain group names with labels (Kiritchenko & Mohammad, 2018; Zhang et al., 2020) or differential performance on data produced by different social groups (Shen et al., 2018; Sap et al., 2019) that causes harm to certain social groups. Several recent studies (Mehrabi et al., 2019; Blodgett et al., 2020; Shah et al., 2020) provide a taxonomy for prediction bias. A number of works propose algorithms to mitigating bias (*e.g.*, Park et al. (2018); Zhang et al. (2018)). We focus on related works attempting to learn de-biased data representations and mitigating bias in pretrained models.

Mitigating Bias in Representations. Towards mitigating bias in models, a popular approach is to interfere with the internal representations of the models. Zhang et al. (2018); Beutel et al. (2017) jointly train a classifier with sensitive attribute predictors with shared representations within an adversarial learning framework. Madras et al. (2018); Elazar & Goldberg (2018) further study the task of learning re-usable de-biased data representations by removing sensitive attributes from data representations. This way, new downstream classifiers (potentially with a different classification task) trained over these de-biased data representations, would not rely on sensitive attributes for prediction. However, because only frozen *data* representations are transferred (instead of the model), the framework cannot generate (de-biased) predictions for new data (*e.g.*, new examples from the same of different domains).

Mitigating bias in Pretrained Models. Another line of work analyzes bias in pretrained models (Zhou et al., 2019; May et al., 2019; Bhardwaj et al., 2020) (*e.g.*, word vectors, BERT). Most of these study measures bias in the representation space only (e.g. representational distances between gender and stereotypes). A few study the propagation of bias to downstream classifiers: Zhao et al. (2019) show models trained over contextualized word embeddings are more gender-biased in coreference resolution tasks (*e.g.*, by associating gendered pronouns with stereotypical words); Liang et al. (2020); Ravfogel et al. (2020) study algorithms to mitigate bias in representations by these pretrained models. However, these work keeps the pretrained model frozen. The key unanswered question with these works is whether mitigating bias in PTLMs leads to downstream models that are similarly less biased in the fine-tuning process.

Unlike the previous line of works, our work does not analyze or mitigate representational bias that originally exists in PTLMs; rather, we study whether by fine-tuning from less biased upstream models benefit downstream classifiers so that they are less biased (potentially by learning to pick up more essential clues (instead of bias) from the data). A few existing works study related research problems, with important differences to our work: Schumann et al. (2019) studies this problem; however, they assume data from the source and the target domain are available at the same time, which is a dissimilar setting to fine-tuning; Shafahi et al. (2020) study transfer learning of adversarial robustness and is related to our study in terms of the methodology, but does not study bias mitigation. In summary, no previous works study the capability of learning less biased downstream classifiers by transferring from a less bias upstream model.

3 Background

We first introduce important concepts, such as transfer learning, prediction bias, and bias mitigation algorithms. We also introduce our problem formulation in this section.

3.1 TRANSFER LEARNING VIA FINE-TUNING

Fine-tuning is a practice of transfer learning which has become common in NLP. A significant advantage is the simplicity of fine-tuning: model developers simply load pretrained weights and take advantage of powerful inductive bias in the source model to obtain a powerful model without the burden of training them from scratch (Qiu et al., 2020).

Formally, we assume a model $f : \mathbb{R}^m \rightarrow \{0, 1, ..., C\}$ maps an input sentence $\mathbf{x}$ to a class label y. The model $f = g \circ h$ is composed by a text encoder $g : \mathbb{R}^m \rightarrow \mathbb{R}^d$ (*e.g.*, Transformers in RoBERTa (Liu et al., 2019)) which maps an input sentence $\mathbf{x}$ to a hidden representation $\mathbf{z} \in \mathbb{R}^d$ in the representation space; and a classifier head $h : \mathbb{R}^d \rightarrow \{0, 1, ..., C\}$, which maps the representation $\mathbf{z}$ to the label space. The (upstream) source model $f_s = g_s \circ h_s$, is first trained in a source domain $\mathcal{D}_s$. Then, for transfer learning in a target domain $\mathcal{D}_t$, the encoder g_t in the target (downstream) model $f_t = g_t \circ h_t$ is initialized as g_s, while the classifier head h_t is randomly initialized. The full target model f_t is trained on the target domain $\mathcal{D}_t$.

3.2 PREDICTION BIAS AND BIAS MITIGATION

Definition of Prediction Bias. Our definition of prediction bias in models follows *equal opportunities* in previous literature (Hardt et al., 2016): a model f is biased if it yields a higher prediction error rates on a subset S over the rest for a certain label, where S typically refers to a sub-population that is either mentioned or authors the data in question. There can be multiple bias factors, indexed with $j \in \{0, 1, ..., K\}$: for example, the subset $S^{(j)}$ may consist of all examples produced by African American English speakers, or all input examples that contain a certain group identifier mention (*e.g.*, gendered pronouns); we use a binary attribute indicator $\mathbf{a}_i \in \{0, 1\}^K$, where each item $a_i^{(j)} \in \{0, 1\}$ indicates whether an example $(\mathbf{x}_i, y_i)$ belongs to the set $S^{(j)}$. We focus on false positive error rates for disadvantaged outcomes (*e.g.*, sentences misclassified as spam or offensive) in this paper: in all tasks that we experiment with, we consider false positives to be more harmful; however, the framework also applies to other bias, *e.g.*, where false negative rates are equally important. For simplicity of notations, we use label $y = 1$ to refer to disadvantaged outcomes, and $y = 0$ for others. The bias is formally quantified as the false positive rate differences (FPRD), where,

$$\text{FPRD}^{(j)} = P(\hat{y} = 1 | a^{(j)} = 1, y = 0) - P(\hat{y} = 1 | a^{(j)} = 0, y = 0) \tag{1}$$

where $\hat{y}$ is the model prediction. The term $P(\hat{y} = 1 | y = 0, a^{(j)})$ is the false positive rate (FPR), *i.e.*, the probability of predicting a label as positive when the ground truth is negative, conditioned on the value of the attribute $a^{(j)}$.

Bias Mitigation Approaches. Bias mitigation algorithms have been studied to reduce bias in models. Given a labeled datasets with attribute annotations $\{(\mathbf{x}_i, y_i, \mathbf{a}_i)\}_{i=1}^N$, the model is trained by jointly optimizing a main learning objective ℓ_c with a bias mitigation objective $\ell_b(\mathbf{x}, y, \mathbf{a})$, which penalizes biased behaviors of models. For example, in adversarial learning, the loss terms coerce attribute a to be predictable from the data representation $\mathbf{z} = g(\mathbf{x})$. Note that while $\mathbf{x}$ or y can be excluded for computing the loss ℓ_b, the ground truth attribute indicator $\mathbf{a}$ is almost always required.

Bias mitigation algorithms face two major challenges in practice: (1) Attribute annotations $\mathbf{a}_i$ are usually not provided in datasets (e.g., we might not have access to social media users' demographics) and may be difficult to collect due to privacy concerns. (2) The bias mitigation process is applied repetitively to *each* downstream application, resulting in increased running time and human labor.

3.3 PROBLEM FORMULATION

The limitation of de-biasing algorithms motivates us to set up a new problem formulation that separates out the de-biasing phase from downstream classifier training. Formally, we consider the problem of reducing bias in (potentially a large number of) downstream classifiers $f_t = g_t \circ h_t$ in target domains $\mathcal{D}_t$ with a training set $\{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^N$. The attribute annotations $\mathbf{a}_i$ are not available in the target domain. The target model f_t can be fine-tuned from a model f_s trained in a source domain $\{(\mathbf{x}_i, y_i, \mathbf{a}_i)\}_{i=1}^N$, where it capable of running bias mitigation algorithms. We further consider two dimensions of settings, divided by: (1) the similarity between the source and the target domain, and (2) the number of bias factors considered. We illustrates different settings in Figure 1 (left).

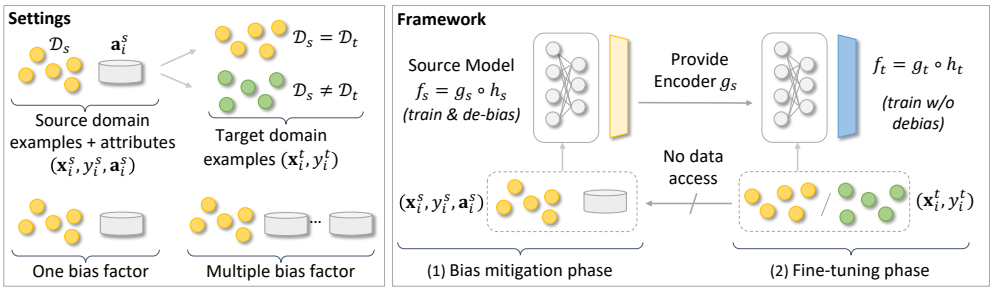

Figure 1: Proposed settings and the framework of upstream bias mitigation in transfer learning.

Same or Different Source and Target Domains. In the simplest setting, we have $\mathcal{D}_s = \mathcal{D}_t$. Practically, it corresponds to training a de-biased model over examples from the same domain. In a more challenging setting, we have $\mathcal{D}_s \neq \mathcal{D}_t$, which is a more popular case in practice.

Dealing with One or Multiple Bias Factors. The source model f_s can be de-biased for either one or multiple bias factors (*e.g.*, both the gender bias and the racial bias).

4 Upstream Bias Mitigation in Transfer Learning

In this section we illustrate our framework of upstream bias mitigation in transfer learning. The framework involves two phases: **(1) Bias Mitigation phase** (Sec. 4.1). a source model $f_s = g_s \circ h_s$ is trained on the source data, with bias mitigation objectives; then, the text encoder g_s (*i.e.*, the classifier head h_s is discarded) is provided to a target domain. **(2) Fine-tuning phase** (Sec. 4.2). The target model $f_t = g_t \circ h_t$ use g_s as weight initialization and fine-tune for the in-domain prediction performance only in the target domain. Figure 1 (right) illustrates the proposed framework.

4.1 Bias Mitigation Algorithms

We consider two approaches for model de-biasing in the de-biasing phase: explanation regularization (Kennedy et al., 2020) and adversarial de-biasing (Elazar & Goldberg, 2018; Xia et al., 2020). The explanation regularization approach applies when there are a set of lexicons that models may be overly sensitive to: for example, the model may be spuriously correlate labels with group identifier mentions, or terms frequently used by a certain demographic group in the input sentence, which may result in higher FPRD on examples that contain these lexicons. The adversarial de-biasing approach is more general and does not require a predefined set of lexicons.

Explanation Regularization. Explanation regularization approaches assign an importance score $\phi(w, \mathbf{x})$ for each word w in a predefined lexicon set $\mathcal{S}$ that is present in the input example $\mathbf{x}$, and regularize the importance score jointly with the main learning objective. The jointly learning objective is written as,

$$\min_f \quad \ell_c + \alpha||\phi(w, \mathbf{x})||^2 \tag{2}$$

where ℓ_c is the classification loss and $\alpha||\phi(w, \mathbf{x})||^2$ penalizes the importance attributed cumulatively to all $w \in \mathcal{S}$. α is a hyperparameter controlling the strength of the regularization. While a variety of explanation algorithms can be applied here, we focus on the most simple input occlusion algorithm (Zintgraf et al., 2017), where the importance is measured as the model prediction change when the term w is removed from the input $\mathbf{x}$.

Adversarial De-biasing. Adversarial de-biasing algorithms reduce information about the attributes $\mathbf{a}$ encoded in the intermediate representations by the encoder g. During training, an adversarial classifier head (an MLP) $h_i^{adv} : \mathbb{R}^n \rightarrow [0, 1]$ is built upon the encoder g for each attribute $a^{(j)} \in \mathbf{a}$. The classifier is trained to predict the attribute a by optimizing the cross entropy loss $\ell_{adv}(g \circ h_{adv}(\mathbf{x}), a)$. A gradient reversal layer (Ganin et al., 2016) is added between the encoder g and h_{adv}, so that the encoder g is optimized to generate representations that do not encode information about the attribute a. The adversarial learning objective is optimized jointly with the classification

objective at training. Formally, the optimization problem is written as,

$$\min_{g,h} \max_{h^{adv}_{1:K}} \quad \ell_c + \sum_{j=1}^{K} \ell_{adv}(g \circ h^{adv}_j(\mathbf{x}), a_j) \tag{3}$$

while the adversarial loss ℓ_{adv} can take other forms as in Madras et al. (2018), we did not observe significant differences empirically.

We train a de-biased classifier with either explanation regularization or adversarial de-biasing in the source domain. The text encoder g_s is then extracted from the full model $f_s = g_s \circ h_s$ provided to the target domain to perform fine-tuning.

4.2 Model Fine-tuning Algorithms

In the second phase, we train a downstream classifier $f_t = g_t \circ h_t$ in the target domain $\mathcal{D}_t$ without de-biasing algorithms, where g_t is initialized with g_s from the source model. We assume that the full-model is fine-tuned, *i.e.*, both g_t and h_t are trained jointly. We additionally consider two approaches: (1) the encoder g_t is frozen while only h_t is trained, and when (2) the g_t is penalized for deviating from g_s with a regularization term. We use the ℓ^2-sp regularizer (Li et al., 2018), which penalizes the distance between the weights and the initial point of fine-tuning. Formally, let $\mathbf{w}_0$ be the initial weight of the encoder g before fine-tuning, and $\mathbf{w}$ be the current weight of g. The ℓ^2-sp regularizer is written as $\Omega(\mathbf{w}) = \beta||\mathbf{w} - \mathbf{w}_0||^2_2$, where β is a hyperparameter controlling the strength of the regularization, set to 1 by default.

5 Experiments

In this section we describe the general experimental framework and in each sub-section we report each variation of this setup. We consider two bias factors in our study, namely the *group identifier bias* and the *African American English (AAE) dialect bias*.

Group Identifier Bias in Hate Speech Detection. The bias refers to sentences containing group identifiers that are more likely to be misclassified as hate speech over others. This behavior is harmful to certain demographic groups by misclassifying innocuous text with the group identifier mentions (*e.g.*, I am a muslim) as the hate speech. We include two datasets for study, namely the Gab Hate Corpus (GHC) (Kennedy et al., 2018) and the Stormfront corpus (de Gibert et al., 2018). Both datasets contain binary labels for hate and non-hate instances. We use the explanation regularization approach with the 25 group identifier lexicons provided in (Kennedy et al., 2020).

AAE Dialect Bias. Sap et al. (2019) show offensive and hate speech classifiers yield a higher false positive rate on the text written in African American English (AAE). This bias brings significant harm to the community that uses AAE, for example, by leading to the disproportionate removal of the text written AAE (that is presumed to be hateful or offensive) in social media platforms, and additionally reinforces negative perceptions of AAE. We include two datasets for study: FDCL (Founta et al., 2018), which is a four-way classification dataset among *normal*, *abusive*, *hateful*, and *spam*; DWMW (Davidson et al., 2017), which is a three-way classification dataset among *normal*, *abusive*, and *hateful*. The models are trained to perform four-way and three-way classification respectively; while to evaluate the prediction bias, we treat *abusive*, *hateful* and *spam* together as disadvantaged outcomes. We use an off-the-shelf AAE dialect predictor (Blodgett et al., 2016) to identify examples written in AAE for the bias mitigation phase. We report the results of both explanation regularization and adversarial learning. See appendix for details of regularization methods.

Metrics. We expect a model to be unbiased while maintaining in-domain classification performance. To evaluate the in-domain classification performance, we report F1-scores for GHC and Stormfront, and the accuracy scores for FDCL and DWMW. Following Zhang et al. (2018), we use the equal error rate (EER) threshold for prediction, *i.e.*, we set the prediction threshold so that the overall false positive rates and the false negative rates are the same on the validation set.

To evaluate the group identifier bias on GHC and Stormfront, the FPRD metrics defined in Eq. 1 can be directly applied on the in-domain test set, where the subset S consists of examples containing one of 25 group identifiers provided by Kennedy et al. (2020). In addition, in order to experiment on examples that contained group identifiers but were not hate speech, we followed Kennedy et al.

Method / Datasets	GHC				Stormfront			
Metrics	In-domain F1 (↑)	In-domain FPRD (↓)	IPTTS FPRD (↓)	NYT Acc (↑)	In-domain F1 (↑)	In-domain FPRD (↓)	IPTTS FPRD (↓)	NYT Acc (↑)
	Single dataset				*Single dataset*			
Vanilla	**49.60 ± 1.0**	46.43 ± 2.5	20.01 ± 5.7	72.08 ± 7.3	**53.74 ± 2.8**	18.09 ± 2.7	11.51 ± 5.1	73.06 ± 10
Expl. Reg.	43.37 ± 1.8	**29.29 ± 1.2**	**4.2 ± 1.6**	**81.22 ± 11**	51.53 ± 1.1	**13.43 ± 1.5**	**3.8 ± 0.4**	**83.73 ± 8.0**
	Stf. → GHC				*GHC → Stf.*			
Van. Transfer	47.83 ± 2.1	47.51 ± 4.6	18.03 ± 3.9	66.71 ± 11	55.79 ± 1.3	17.83 ± 2.2	7.27 ± 1.7	76.98 ± 1.1
Expl. Reg. Transfer	**49.94 ± 1.0**	**42.71 ± 3.8**	**12.23 ± 3.3**	**75.34 ± 4.8**	**56.43 ± 0.6**	**18.03 ± 2.5**	**6.86 ± 1.1**	**81.18 ± 1.1**
	Stf. + FDCL → GHC				*GHC+ FDCL → Stf.*			
Van. + Van. Trans.	49.71 ± 0.3	**45.84 ± 3.8**	12.43 ± 2.5	**72.37 ± 7.4**	**56.78 ± 1.6**	**14.26 ± 0.8**	11.04 ± 0.7	77.06 ± 5.1
Expl. Reg + Expl. Reg-Trans.	**50.21 ± 1.4**	47.63 ± 0.7	**12.29 ± 2.7**	68.44 ± 8.6	53.87 ± 1.2	15.92 ± 1.2	**8.4 ± 1.4**	83.71 ± 3.2
Expl. Reg + Adv-Trans.	49.89 ± 1.7	47.85 ± 1.2	21.25 ± 2.0	65.78 ± 5.7	53.63 ± 0.7	15.52 ± 2.2	8.9 ± 1.6	**84.87 ± 1.1**

Table 1: **Cross-domain** transfer learning of less biased models with GHC and Stormfront as target domains. In-domain FPRD, IPTTS FPRD and NYT Accuracy (Acc.) measures prediction bias. The preferred outcomes for each metric are marked with arrows.

(2020) in using a corpus of New York Times articles (NYT), which is a collection of $12.5k$ all non-hate sentences from New York Times articles where each sentence include one of 25 group identifiers. This corpus specifically provides an opportunity to measure FPR, given that models are often biased towards false positives when one of 25 group identifiers are present. We report the accuracy (NYT Acc.) *i.e.*, 1-FPR. Additionally, following the evaluation protocol of Dixon et al. (2018) and Zhang et al. (2020), we also incorporate the Identity Phrase Templates Test Sets (IPTTS), which consists of $77k$ hate and non-hate examples with group identifier mentions generated with templates[2].

To evaluate the AAE dialect bias on FDCL and DWMW, given that only a small number of examples in the datasets are written in AAE and these labels themselves are noisy outputs from a classifier, the in-domain FPRD metrics can be very noisy. Therefore, following previous study, we incorporate the BROD (Blodgett et al., 2016) dataset, which is a large unlabeled collection of twitter posts written by AAE speakers. We sample $20k$ examples with the confidence of AAE speaker larger than 80%. Following Xia et al. (2020), considering that hateful, abusive and spam posts only take a small portion, we treat all twitter posts from the BROD dataset as the label *normal*, and report the accuracy (which equals 1-FPR) on the dataset, noted as BROD Acc.

Compared Methods. For comparison, we report the results when (1) the downstream classifier is directly trained without any regularization (noted as **Vanilla**); and (2) the downstream classifier is trained in the target domain with explanation regularization or adversarial de-biasing (noted as **Expl. Reg.** or **Adv. Learning**). We also report (3) a model fine-tuned from a naively trained model in the source domain(noted as **Van-Transfer**); and finally, we report the results of the proposed framework, where the model is fine-tuned from a less biased model in the source domain (noted as **Expl. Reg-Transfer** or **Adv-Transfer**). The Van-Transfer serves as a baseline and dissects the benefit of bias mitigation performed in the source domain. We expect that our framework achieves a similar in-domain performance (In-domain F1 or In-domain Accuracy), and less bias (lower FPRD in the in-domain test sets and IPTTS; higher accuracy on NYT and BROD) compared to Vanilla and Van-Transfer. Note that we do not expect our framework to achieve a lower bias compared to directly mitigating bias in target domain; the focus of our study is to reduce bias when bias mitigation is not practical in the target domain, as discussed in Sec. 3.2.

Implementation Details. We use RoBERTa-base as our text encoder (*i.e.*, the model in the bias mitigation phase itself is fine-tuned from RoBERTa-base). See Appendix for more details.

5.1 Cross-domain Transfer Learning of Less Biased Models

As one of the main part in our study, we first show the results when the source and the target domains are different. For notations, we show the source and the target domains in the left and the right-hand side of the arrow respectively. We perform transfer learning from GHC to Stormfront (GHC → Stf.), from Stormfront to GHC (Stf. → GHC); and perform transfer learning from FDCL to DWMW (FDCL → DWMW). We do not use DWMW as the source domain: as a great portion of offensive text are in AAE for the dataset, bias mitigation algorithms make a marginal effect even when it is directly applied on the dataset — which is also observed by Xia et al. (2020).

[2]Following these works, for IPTTS we compute FPRD as $\sum_z |\mathrm{FPR}_z - \mathrm{FPR}_{overall}|$, where FPR_z is false positive rate on sentences with the group identifier z, and $\mathrm{FPR}_{overall}$ is the overall FPR.

Method / Datasets	DWMW	
Metrics	In-domain Acc. (↑)	BROD Acc. (↑)
	Single dataset	
Vanilla	**91.46 ± 0.1**	**78.77 ± 0.3**
Expl. Reg.	91.38 ± 0.1	76.61 ± 1.5
Adv. Learning	91.11 ± 0.3	77.53 ± 0.9
	FDCL → DWMW	
Van. Transfer	91.27 ± 0.2	78.98 ± 1.1
Expl. Reg. Transfer	91.39 ± 0.0	**80.27 ± 0.2**
Adv. Transfer	**91.60 ± 0.1**	79.98 ± 1.7
	GHC + FDCL → DWMW	
Van. + Van.	91.65 ± 0.1	80.98 ± 0.4
Expl. Reg + Expl. Reg.	**91.79 ± 0.4**	**81.36 ± 0.8**
Expl. Reg + Adv.	91.33 ± 0.1	81.09 ± 0.4
	Stf. + FDCL → DWMW	
Van. + Van.	91.64 ± 0.2	81.12 ± 0.1
Expl. Reg + Expl. Reg.	**91.66 ± 0.2**	80.05 ± 0.1
Expl. Reg + Adv.	91.55 ± 0.2	**81.14 ± 1.5**

Table 2: **Cross-domain** transfer with DWMW as the target domain.

Method / Datasets	FDCL (Half-)B	
Metrics	In-domain Acc. (↑)	BROD Acc. (↑)
	Single dataset	
Vanilla	75.72 ± 0.2	73.57 ± 1.2
Expl. Reg.	**77.30 ± 0.2**	76.72 ± 1.1
Adv. Learning	75.28 ± 0.2	**77.12 ± 1.2**
	FDCL A → FDCL B	
Van. Transfer	**76.33 ± 0.6**	70.35 ± 2.4
Expl. Reg. Transfer	76.22 ± 0.5	69.29 ± 1.8
Adv. Transfer	75.88 ± 0.4	**71.11 ± 1.6**
	GHC A + FDCL A → FDCL B	
Van. + Van.	**77.34 ± 0.4**	72.96 ± 1.5
Expl. Reg + Expl. Reg.	76.21 ± 0.4	73.10 ± 1.4
Expl. Reg + Adv.	76.94 ± 0.4	**76.55 ± 0.7**
	Stf. A + FDCL A → FDCL B	
Van. + Van.	**77.18 ± 0.5**	71.12 ± 1.2
Expl. Reg + Expl. Reg.	77.13 ± 0.3	72.17 ± 1.6
Expl. Reg + Adv.	76.64 ± 0.6	**76.55 ± 0.6**

Table 3: **Same-domain** transfer with the second half of FDCL as the target domain.

Tables 1 and 2 show the results of the cross-domain transfer between GHC and Stormfront and from FDCL to DWMW. We summarize our findings below.

Reduced bias compared to vanilla training on a single dataset. From the results of cross-domain transfer learning when only one dataset is included in the source domain (*i.e.*, Stf.→Gab, Gab→Stf., FDCL→DWMW), we see reduced bias in the target domain by transferring from a less biased model compared to Vanilla: FPRD on IPTTS has decreased from 20.01 to 12.23 for GHC and from 11.51 to 6.86 for Stormfront; Accuracy on NYT has increased from 72.08 to 75.34 for GHC, and from 73.06 to 81.18 for Stormfront; Accuracy on BROD has increased from 78.77 to 80.27 (Expl. Reg-Transfer) for DWMW. Meanwhile, the in-domain prediction accuracy has also increased (on GHC and Stormfront) or preserved (on DWMW). It is notable that on directly running bias mitigation algorithms on DWMW is not effective; while transferring from FDCL improves BROD Acc. The In-domain FPRD on Stromfront is an only exception: however, as discussed in our metrics section, the in-domain FPRD is computed over a much smaller set of examples compared to NYT and IPTTS, which make the score less reliable.

Transfer learning itself makes a positive impact, while mitigating bias the source model further reduces bias. We notice transfer learning itself has an overall positive impact on reducing bias from the comparison between Vanilla and Van-Transfer. The observation aligns with Tu et al. (2020), where the authors show multi-task learning improves models' robustness to spurious correlations. We have shown the conclusion extends to the transfer learning case in bias mitigation. Nevertheless, transfer learning from a less biased model almost always further reduce the bias.

5.2 Dealing with Multiple Bias Factors

We also show the results where we include more than one datasets and simultaneously remove both the group identifier bias and the AAE dialect bias in the source domain with two different datasets. We apply multi-task learning (MTL), where two datasets share the same text encoder g but use different classification heads h_t^1 and h_t^2 and trained jointly. We train the source model with the combination of GHC and FDCL (GHC+FDCL) or Stormfront and FDCL (Stf. + FDCL), and fine-tune the model on Gab, Stormfront, or DWMW. The results are shown in Table 1 under the rows Stf. + FDCL → GHC and GHC + FDCL → Stf., and in Table 2 under the rows Stf. + FDCL → DWMW, GHC + FDCL → DWMW. We summarize the observations below.

Transfer learning from less biased models on GHC + FDCL reduces both kinds of bias. We find fine-tuning from Expl. Reg + Adv. and Expl. Reg + Expl. Reg (trained on GHC+FDCL) reduces group identifier bias in Stormfront in FDCL, compared to both Vanilla and Van. + Van-Transfer (measured with IPTTS and NYT). On DWMW, we see that transfer learning itself (Van.+Van-Transfer) has already improved the accuracy score on BROD from 79.77 to 80.98, and Expl. Reg + Adv-Transfer and Expl. Reg + Expl. Reg-Transfer has slightly improved the accuracy further to 81.09 and 81.36 respectively.

Transferring from less biased models on Stf. + FDCL does not further improve over Van.+Van-Transfer. Similarly, fine-tuning from vanilla models trained on Stormfront and FDCL improves bias metrics compared to vanilla training on GHC or FDCL. However, we see mitigating bias in source models does not bring further improvements. The results imply different tasks may interfere with each other when applied for mitigating bias in upstream models. We leave further study and solution to such interference as a future work.

5.3 Transfer Learning with the same Source and Target Domains

We further show the results when the target domain consists of new emerging examples from the same data distribution as the source domain, which is a simpler setting than the cross-domain setup. The experiments allows us to solely focus on whether learned bias mitigation in the source domain is preserved in fine tuning and discard the challenge of domain dissimilarity. We set up the experiments by partitioning GHC, Stormfront and FDCL to two subsets with equal sizes, noted as subsets A and B of corresponding datasets. Subsets A and B are regarded as the source domain and the target domain respectively. Similar to the cross domain setting, we show the results when the models are only trained in the target domain, when a single bias factor is reduced from the source model, and when two bias factors are reduced.

The similarity between the source and the target domain enables better transfer learning of less biased models. Table 3 resent the results on FDCL. We also include the results for GHC and Stormfront in Appendix (Table 5). By comparing between Van-Transfer or Van.+Van-Transfer and transfer learning from regularized models, we see a clear effect of mitigating bias in the source model compared to the cross-domain transfer learning setup.

In the experiments above, we have shown that mitigating bias in the source model provides inductive bias to target model that is preserved in the fine-tuning. Next, we study whether freezing the weights or discouraging the weight change allows better mitigation of bias in the target domain.

5.4 Freezing or Regularizing Model Weights

Intuitively, freezing or discouraging the changes on the encoders may help to retain the knowledge in the encoder and reduce bias in the target domain. However, we show a counter-intuitive conclusion: most of times freezing or discouraging weight changes does not contribute to bias mitigation (while also decreasing the in-domain classification performance). Table 4 show the results when we keep the weights frozen (Freeze), regularized from changing (ℓ^2-sp), or unregularized at fine-tuning (fine-tune). In Stf. $\rightarrow$ GHC, freezing the weights contributed to reducing the bias, while ℓ^2-sp failed to help. We tried also other regularization strengths β but did not observe better results. In GHC $\rightarrow$ Stf and FDCL $\rightarrow$ DWMW, both freezing the weight and ℓ^2-sp increases the bias. A possible reason is that when the encoder is frozen, we are training a linear model; simple models are known to more biased because they can only make use of superficial clues in the inputs that may spuriously correlate with the labels.

Metrics	In-domain F1 (↑)	In-domain FPRD (↓)	IPTTS FPRD (↓)	NYT Acc. (↑)
	Stf. → GHC			
Freeze	45.42	**37.71**	**7.82**	**84.45**
ℓ^2**-sp**	49.31	47.03	14.24	71.88
Fine-tune	**49.94**	42.71	12.23	75.34
	GHC → Stf.			
Freeze	47.32	25.02	8.24	64.60
ℓ^2**-sp**	55.80	19.75	6.72	80.42
Fine-tune	**56.43**	**18.03**	**6.86**	**81.88**

(a) GHC to Stf. and Stf. to GHC

Metrics	In-domain Accuracy. (↑)	BROD Accuracy (↑)
	FDCL → DWMW	
Freeze	83.25	64.80
ℓ^2**-sp**	91.38	79.95
Fine-tune	**91.60**	**79.98**

(b) FDCL to DWMW

Table 4: Transferring from explanation regularized models (GHC, Stf) or adversarially debiased models (FDCL) while keeping the encoder frozen (Freeze), fine-tuning with ℓ^2-sp regularizer, or vanilla fine-tuning (Fine-tune).

6 Conclusions

In this paper, we proposed a transfer learning framework for learning less biased downstream models. We experimented with various settings, when single or multiple attributes are included for de-biasing, and when the source and the target domains are the same or different, and obtain overall positive results. For future works, we would study algorithms to learn more transferable less biased upstream models.

REFERENCES

Alex Beutel, J. Chen, Zhe Zhao, and Ed Huai hsin Chi. Data decisions and theoretical implications when adversarially learning fair representations. *ArXiv*, abs/1707.00075, 2017.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in bert. *ArXiv*, abs/2009.05021, 2020.

Su Lin Blodgett, L. Green, and Brendan T. O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *EMNLP*, 2016.

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

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *ArXiv*, abs/1607.06520, 2016.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *ICWSM*, 2017.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

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

Yanai Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. *ArXiv*, abs/1808.06640, 2018.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, I. Leontiadis, Jeremy Blackburn, G. Stringhini, Athena Vakali, M. Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *ICWSM*, 2018.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

M. Hardt, E. Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

B. Kennedy, Xisen Jin, Aida Mostafazadeh Davani, M. Dehghani, and X. Ren. Contextualizing hate speech classifiers with post-hoc explanation. *ArXiv*, abs/2005.02439, 2020.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18, 2018.

Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In **SEM@NAACL-HLT*, 2018.

Keita Kurita, N. Vyas, Ayush Pareek, A. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *ArXiv*, abs/1906.07337, 2019.

Xuhong Li, Yves Grandvalet, and F. Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, 2018.

P. P. Liang, I. Li, E. Zheng, Yao Chong Lim, R. Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. *ACL*, 2020.

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

David Madras, Elliot Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. *NAACL-HLT*, 2019.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ArXiv*, abs/1908.09635, 2019.

J. Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *EMNLP*, 2018.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, N. Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *ArXiv*, abs/2003.08271, 2020.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *ACL*, 2020.

Maarten Sap, D. Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *ACL*, 2019.

Candice Schumann, Xuezhi Wang, Alex Beutel, J. Chen, Hai Qian, and Ed Huai hsin Chi. Transfer of machine learning fairness across domains. *ArXiv*, abs/1906.09688, 2019.

A. Shafahi, Parsa Saadatpanah, C. Zhu, Amin Ghiasi, C. Studer, D. Jacobs, and T. Goldstein. Adversarially robust transfer learning. *ICLR*, 2020.

Deven Shah, H. A. Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In *ACL*, 2020.

Judy Hanwen Shen, Lauren Fratamico, I. Rahwan, and Alexander M. Rush. Darling or babygirl? investigating stylistic bias in sentiment analysis. 2018.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *ArXiv*, abs/2007.06778, 2020.

M. Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. *ArXiv*, abs/2005.12246, 2020.

B. H. Zhang, B. Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *ACL*, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, V. Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *NAACL-HLT*, 2019.

P. Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender. In *EMNLP/IJCNLP*, 2019.

Luisa M. Zintgraf, T. Cohen, Tameem Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. *ArXiv*, abs/1702.04595, 2017.

Method / Datasets	GHC (Half-)B				Stormfront (Half-)B			
Metrics	**In-domain F1** (↑)	**In-domain FPRD** (↓)	**IPTTS FPRD** (↓)	**NYT Acc** (↑)	**In-domain F1** (↑)	**In-domain FPRD** (↓)	**IPTTS FPRD** (↓)	**NYT Acc** (↑)
	Single dataset				*Single dataset*			
Vanilla	37.91 ± 2.5	36.54 ± 2.2	21.50 ± 2.8	68.55 ± 20	**55.56 ± 0.5**	20.81 ± 4.9	10.99 ± 5.6	**66.28 ± 8.1**
Expl. Reg.	**38.09 ± 2.7**	**18.68 ± 0.3**	**4.82 ± 1.1**	**84.05 ± 3.0**	53.05 ± 1.0	**15.97 ± 1.1**	**3.36 ± 3.3**	65.23 ± 10
	GHC A → GHC B				*Stf. A → Stf. B*			
Van-Transfer	42.41 ± 1.0	37.44 ± 1.5	17.67 ± 2.1	75.35 ± 4.2	58.43 ± 1.2	17.96 ± 3.6	11.58 ± 4.8	**74.12 ± 4.2**
Expl. Reg-Transfer	**43.79 ± 1.9**	**34.34 ± 3.1**	**10.02 ± 1.1**	**81.4 ± 1.4**	**58.56 ± 1.0**	**16.42 ± 0.9**	**7.51 ± 2.4**	69.45 ± 4.0
	GHC A + FDCL A → GHC B				*Stf. A + FDCL A → Stf. B*			
Van. + Van.	**44.30 ± 0.7**	41.06 ± 3.9	19.75 ± 6.9	74.60 ± 6.3	**57.58 ± 2.7**	**13.97 ± 2.0**	11.33 ± 2.2	75.72 ± 7.5
Expl. Reg + Expl. Reg-Trans.	42.96 ± 2.0	33.98 ± 3.0	**9.30 ± 2.1**	**86.05 ± 1.9**	56.72 ± 1.7	17.91 ± 1.0	**8.05 ± 0.6**	**77.40 ± 0.3**
Expl. Reg + Adv-Trans.	42.44 ± 3.5	**33.96 ± 1.5**	16.58 ± 1.7	81.79 ± 9.0	55.63 ± 2.5	17.14 ± 0.5	13.78 ± 4.3	70.37 ± 10

Table 5: **Same-domain** transfer with the second half of GHC and Stormfront as the target domain.

A Implementation Details

Training. In the bias mitigation phase, the models are trained with a learning rate $1e^{-5}$, and the checkpoint with the best validation F1 or accuracy score is provided to the fine-tuning phase. We train Gab, FDCL, and DWMW for maximum 5 epochs and Stormfront for maximum 10 epochs. In the fine-tuning phase, we try the learning rate $1e^{-5}$ and $5e^{-6}$, and report the results with a higher validation F1 or accuracy.

Bias mitigation algorithms. For explanation regularization algorithm, we set the regularization strength α as 0.03 for Gab and Stormfront experiments, and 0.1 for FDCL and DWMW experiments. We regularize importance score on 25 group identifiers in Kennedy et al. (2018) for Gab and Stormfront. These group identifiers the ones that have the largest coefficient in a bag-of-words linear classifier. For FDCL, we extract 50 words with largest coefficient in the bag-of-words linear classifier with a AAE dialect probability higher than 60% (given by the off-the-shelf AAE dialect predictor Blodgett et al. (2016)) on its own. For adversarial de-biasing, the adversarial loss term has the same weight as the classification loss term.

B Complete Analysis of Same-Domain Transfer

Table 5 show the results of same-domain transfer with the second half of GHC and Stormfront datasets as the target domain. Similar to the cross-domain setup, transfer learning from a less biased model overall reduces the bias compared to Vanilla and Vanilla-Transfer. The observation is consistent when multiple bias factors are reduced in the source domain. We find the NYT accuracy on Stormfront is an exception, which is not improved even when we directly run explanation regularization in the target domain. We reason that the Half-Stormfront dataset is small and the average length of the sentences are quite different between Stormfront and NYT, so that a model trained on Stormfront hardly generalizes to NYT.