FairFlow: Mitigating Dataset Biases through Undecided Learning for Natural Language Understanding

Jiali Cheng and **Hadi Amiri**
University of Massachusetts Lowell
{jiali_cheng, hadi_amiri}@uml.edu

Abstract

Language models are prone to dataset biases, known as shortcuts and spurious correlations in data, which often result in performance drop on new data. We present a new debiasing framework called "FairFlow" that mitigates dataset biases by learning to be *undecided* in its predictions for data samples or representations associated with known or unknown biases. The framework introduces two key components: a suite of data and model perturbation operations that generate different biased views of input samples, and a contrastive objective that learns debiased and robust representations from the resulting biased views of samples. Experiments show that FairFlow outperforms existing debiasing methods, particularly against out-of-domain and hard test samples without compromising the in-domain performance[1].

1 Introduction

Existing computational models developed for natural language processing (NLP) tasks are vulnerable to dataset biases and spurious correlations in data, often referred to as "shortcuts." These shortcuts enable models to achieve high performance on NLP datasets by exploiting surface-level correlations between features and labels. However, they also result in a significant performance drop on hard or slightly modified test data (Naik et al., 2018). For example, in the area of natural language inference (NLI), models like BERT (Devlin et al., 2019) tend to misclassify premise-hypothesis pairs that contain "negation" words in their hypotheses as "contradiction," which happen to be predictive features associated with the *contradiction* label in certain NLI datasets (Gururangan et al., 2018; Poliak et al., 2018; Modarressi et al., 2023).

Existing debiasing approaches can detect known (Clark et al., 2019; Sanh et al., 2021; Karimi Mahabadi et al., 2020; Modarressi et al., 2023) and previously unidentified or unknown (Utama et al., 2020b; Sanh et al., 2021) biases in training data. They mitigate dataset biases by re-weighting examples (Sanh et al., 2021; Karimi Mahabadi et al., 2020), learning robust representations (Gao et al., 2022; Du et al., 2023), learning robust feature interaction patterns (Wang et al., 2023), or reducing the effect of biased model components (Meissner et al., 2022).

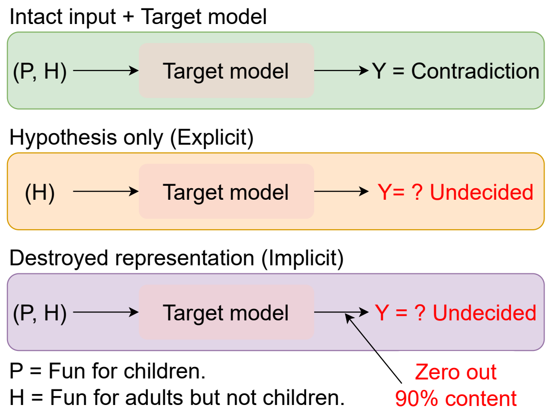

Figure 1: An example highlighting the concept of "undecided learning" using two types of data perturbation techniques. Given a premise-hypothesis pair in NLI, the model is expected to correctly classify their entailment relationship. However, given only the hypothesis, a robust model should be undecided, i.e., refrain from making a definite judgment about the relationship between an unknown premise and the given hypothesis. Similarly, given a severely corrupted representation, a robust model should be undecided about the relation between a corrupted premise and hypothesis pair. Models that retain confidence in assigning labels to such inputs are likely to rely on shortcuts. FairFlow takes an undecided stance against such inputs.

Despite the significant progress made in addressing dataset biases, existing models have certain limitations: **(a)**: they often adopt a *single view* to dataset biases and primarily focus on specific types of biases (Clark et al., 2019; Karimi Mahabadi et al., 2020). However, rich sources and diverse types of dataset biases can be present in the data.

[1]Our code is available at https://github.com/CLU-UML/FairFlow.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21960–21975
November 12-16, 2024 ©2024 Association for Computational Linguistics

(b): existing approaches that are based on weak learners (Utama et al., 2020b; Sanh et al., 2021; Ghaddar et al., 2021; Meissner et al., 2022) rely on a *single* weak learner to identify biases, which inevitably tie their performance to the capabilities of the chosen weak learner. **(c)**: prior works often evaluate debiasing methods using BERT-based models, which may limit their generalizability to other model architectures.

We tackle the above challenges by developing FAIRFLOW–a multiview contrastive learning framework that mitigates dataset biases by being *undecided* in its prediction for biased views of data samples (see **Figure 1**). Specifically, the proposed method employs several data perturbation operators to generate biased views of intact data samples and integrate them into the training data and learning process. When presented with biased inputs, the model is trained to be undecided about the possible labels by making a uniform prediction across the label set. At the same time, the model is encouraged to be confident about intact inputs, which *often* serve as a reference for unbiased samples. Therefore, the approach encourages learning representations that are more attentive to the true signal of the underlying tasks rather than relying on shortcuts that are specific to certain datasets. In addition, the inherent randomness of the implicit perturbations in FairFlow (§2.4.1) exposes the model to a diverse range of perturbations and prevents it from overfitting to specific types of biases present in the data.

The contributions of this paper are:

- categorization of dataset biases: we categorize prevalent data biases in NLU and model them using data perturbation operations;
- bias mitigation as an "undecided learning" problem: we formulate the bias mitigation problem as an "undecided learning" problem, which encourages reliance on genuine and task-related signals for effective debiasing;
- robust performance on challengng samples: our approach shows robust results on *harder* test data while maintaining strong in-domain performance across several NLU tasks.

The experimental results show that FAIRFLOW obtains substantial improvement over competing models. Specifically, it achieves an average performance gain of 10.2 points on stress test datasets across several NLU tasks while maintaining performance on the original test sets. In addition, models trained using our framework show strong transferability, resulting in an average gain of 3.7 points in transfer testing experiments across different datasets and domains. Furthermore, we show that existing methods can be further improved by incorporating the proposed perturbation operators within their original objectives, resulting in a substantial average improvement of 5.8 points on stress test sets across datasets.

2 Method

2.1 Problem Formulation

We consider a dataset $\mathcal{D} = \{(x_i, y_i)|_{i=1}^n\}$, where x_i is the i-th input consisting of several constituents $x_i = (x_i^1, x_i^2, \ldots, x_i^p)$, $|x_i| = p > 1$, and y_i is the corresponding output for x_i. For example, in case of NLI, $p = 2$ represents premise and hypothesis in each input and y_i reflects the entailment or no-entailment relationship between the input pair. Our goal is to develop a model that is robust against different types of dataset biases in $\mathcal{D}$. We note that the model can be applied to a more general setting where input x_i does not explicitly consist of several constituents, see §2.3.1.

2.2 Overview

We categorize dataset biases as *explicit* and *implicit* biases. Explicit biases are readily discernible and understandable by humans, such as high degree of lexical overlap between the premise and hypothesis in case of NLI. On the other hand, implicit biases are often subtle, indiscernible to humans, and more challenging to detect. For example, any word in input has the potential to act a shortcut, resulting in spurious correlations. We introduce different types of explicit and implicit biases that are *task-independent* and generally applicable to bias mitigation across NLP datasets (§2.3). Given such categorization, we propose a debiasing framework that mitigates dataset biases by learning genuine task-related representations that are attentive to the true signal of the tasks rather than biases and shortcut solutions in datasets. The key novelty of our approach is in imposing a downstream model to adopt an "undecided" ("uncertain") stance in its predictions when presented with biased views of inputs. The framework achieves this goal by assigning a uniform probability across the labels, see **Figure 2**. Specifically, the model regularizes the loss of the target task with a contrastive loss which draws biased predictions closer to a uniform

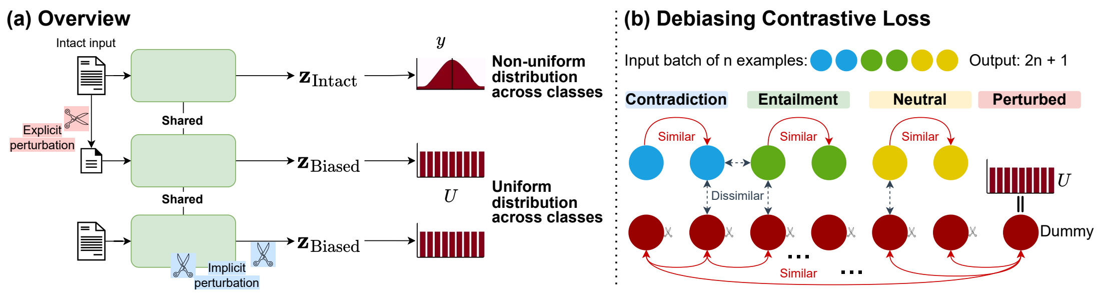

Figure 2: Architecture of the proposed model. (a) Explicit and implicit perturbations are applied to inputs to obtain biased prediction z_{Biased}. (b) Biased predictions are drawn closer to uniform distribution, while predictions for intact input are pushed away from uniform distribution through contrastive learning.

distribution while pushing other predictions away from uniform distribution (§2.4).

2.3 Bias Modeling

We present a series of data perturbation operations to generate biased views by corrupting intact inputs. These perturbations can be explicit or implicit. In explicit perturbation, we directly corrupt the input data, while in implicit perturbation, we corrupt the representations of the input data. These perturbation techniques impose controlled variations on the data, which enable us to conduct a thorough analysis of their effects on bias mitigation.

2.3.1 Explicit Biases

Ungrammatical Perturbation Recently, Sinha et al. (2021) showed that traditional and recent neural language models can be largely invariant to random word order permutation in their inputs. An ungrammatical input is often not understandable by humans and can potentially lead to explicit biases when models confidently predict outcomes for such inputs. For example, a model making a confident prediction about the contradiction class for the following perturbed premise-hypothesis pair from Figure 1 may attribute its confidence to the negation term in the hypothesis: (“`children fun for`”, “`children fun adults but for not`”). To obtain an input with grammatical biases, we design the perturbation operation $\mathcal{P}_{Gra}$ that corrupts the word order in each input x_i. We encode the shuffled input using the shared encoder f and transform it with a branch-specific MLP as follows:

$$z_{Gra} = \mathsf{MLP}_{Gra}\Big(f\big(\mathcal{P}_{Gra}(x_i)\big)\Big). \tag{1}$$

Sub-input Perturbation In NLP tasks that involve multi-part inputs (such as NLI), it is crucial to use the information from all parts of the input for prediction, i.e., all constituents should collectively contribute to accurate prediction. More importantly, an incomplete input should not lead to a confident prediction, as important information may be removed. Therefore, an explicit bias arises when the model makes confident predictions based on incomplete input, such as predicting the *entailment* relation when only the hypothesis is provided as input in case of NLI. Sub-input biases can arise from any part of the input, denoted as $\{x_i^j\}_{j=1}^p$, or from various text spans within different sub-parts. To realize sub-input biases, we define the $\mathcal{P}_{Sub}$ operator that takes one of the constituents of x_i, which is hen encoded with a shared encoder f and further transferred with a constituent-specific Multi-Layer Perceptron MLP_{Sub} as follows:

$$z_{Sub} = \mathsf{MLP}_{Sub}\Big(f\big(\mathcal{P}_{Sub}(x_i)\big)\Big). \tag{2}$$

We note that this operator is applicable to a more general setting where input x_i does not explicitly consist of several constituents, e.g., in general text classification problems. In such cases, each x_i can be divided into $p > 1$ text segments. However, we acknowledge that there are tasks in which one sub-input, i.e. x_i^j for a specific j, is enough to make a correct prediction for the complete input x_i, and therefore remaining undecided may seem counter-intuitive. Nevertheless, by training the model to be undecided when presented with incomplete information, we minimize the risk of biased predictions based solely on partial information, which can, in turn, make the model more robust against potential biases associated with incomplete data.

The idea of implicit perturbations is to obtain biased representations of intact data, without explicitly perturbing the input. We introduce model- and representation-based implicit perturbation.

Model-based Perturbation This approach largely perturbs a given model by converting it into a much weaker model, using mechanisms such as sparsification and layer dropping (Pool and Yu, 2021). A weaker model is believed to capture more biases than a stronger model (Ghaddar et al., 2021; Sanh et al., 2021; Utama et al., 2020b). While existing methods require training a weak learner in advance (Utama et al., 2020b; Sanh et al., 2021; Meissner et al., 2022), our method obtains biased predictions through the same deep neural model (f) and can be trained *end-to-end*. Formally, we design a model-based perturbation operator $\mathcal{P}_{Mod}$ that uses only the first k layers of the shared encoder f, which results in a substantially weakened model with reduced representation power. This branch encodes the intact input using the perturbed model and transform it with a branch-specific MLP as follows:

$$z_{Mod} = \mathsf{MLP}_{Mod}\Big(\mathcal{P}_{Mod}(f)(x_i)\Big). \tag{3}$$

Representation-based Perturbation This perturbation encodes the intact input with the original encoder f but significantly corrupts the generated representations. Given this severely damaged and much less meaningful representation, the model should not be able to predict the correct label. We design a representation-based perturbation operator $\mathcal{P}_{Rep}$ that corrupts the intact representation, $f(x_i)$, and creates a severely perturbed representation. We then transform the perturbed representation with a branch-specific MLP as follows:

$$z_{Rep} = \mathsf{MLP}_{Rep}\Big(\mathcal{P}_{Rep}\big(f(x_i)\big)\Big). \tag{4}$$

Table 1 summarizes the above perturbation operators and provides details of their implementations.

Operator	Type	Implementation
$\mathcal{P}_{Gra}$	Explicit	Shuffle tokens in x_i randomly
$\mathcal{P}_{Sub}$	Explicit	Drop $1/p$ of tokens from x_i^j randomly
$\mathcal{P}_{Sub}$	Explicit	Drop $x_i^j, j = 1 \dots p$
$\mathcal{P}_{Mod}$	Implicit	Use only first k of layers of f
$\mathcal{P}_{Rep}$	Implicit	Zero out $m\%$ of values in $f(x_i)$

Table 1: Implementations of proposed perturbations

2.4 Supervised Contrastive Debiasing

Given the explicit and implicit biased views of data samples, we expect a robust debiasing model to maintain an "undecided" stance across labels for biased inputs while providing confident predictions for intact inputs $x_i, \forall i$. Based on this intuition, the outputs of the bias branches should approximate a *uniform distribution* (U) across classes, while the output of the original branch should align with its corresponding gold distribution, i.e., the label y_i. To achieve this goal, we adapt the supervised contrastive loss (Khosla et al., 2020), which operates by first grouping samples based on their respective labels, and then encouraging predictions (logits) of pairs within the same group to become closer while pushing logits of pairs across different groups further apart, i.e. forming positive pairs within the same group while creating negative pairs using all other pairs:

2.4.1 Implicit Biases

We adapt this loss function for bias mitigation as follows (described for a single perturbation for simplicity): given a batch of n non-perturbed examples, we perturb them using a perturbation technique described in Table 1. The perturbed examples form a single group as they all have the same label (a uniform distribution across all classes), and the non-perturbed examples with the same label form separate groups.[2] As illustrated in **Figure 2**, we encourage the model to be undecided about the label of perturbed inputs by adding a dummy example that has a "fixed" uniform distribution across all labels to the group of perturbed examples, resulting in a batch of $2n+1$ examples ($\mathcal{I}$). We compute the contrastive loss as follows:

$$\mathcal{L}_{\text{Debias}} = \sum_{i\in\mathcal{I}} \frac{-1}{|\mathcal{G}(i)|} \sum_{j\in\mathcal{G}(i)} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k\in\mathcal{A}(i)} \exp(z_i \cdot z_k/\tau)}, \tag{5}$$

where $\mathcal{G}(i)$ is the set of examples that are in the same group as i (having the same label as i); $\mathcal{A}(i) = \mathcal{I}\backslash\{i\}$ is the set of all examples except i; z indicates the logit of an example, which for perturbed examples is obtained from one of the Equations (2)–(4); and τ denotes the temperature parameter.[3] The dummy example in the perturbed group has a fixed uniform distribution across all labels as its z. This formulation encourages the model to be undecided about the label of perturbed inputs, while being confident about the labels of intact inputs, allowing it to effectively distinguish between different groups of examples.

[2] For example, four groups in case of NLI: perturbed examples, non-perturbed examples labeled as 'entailment', non-perturbed examples labeled as 'contradiction', and non-perturbed examples labeled as 'neutral'.

[3] We note that the summation over all samples except i in the denominator of (5) is motivated by noise contrastive estimation and N-pair losses (Khosla et al., 2020; Gutmann and Hyvärinen, 2010; Sohn, 2016), in which the ability to discriminate between signal and noise (negative class) is improved by adding more examples of negative class.

Finally the model learns the debiasing task in an *end-to-end* manner by minimizing the standard cross-entropy loss with predictions of intact input $z_{\text{Intact}} = f(x_i)$ and the debiasing loss, weighted by a balancing hyperparameter λ as follows:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{CE}}(z_{\text{Intact}}, y_i) + \lambda \mathcal{L}_{\text{Debias}}. \quad (6)$$

Compatibility and Difference with Other Debiasing Objectives and Training Methods Our framework is designed to be compatible with debiasing objectives in existing literature. Notably, it can incorporate objectives such as the product of experts (PoE) (Karimi Mahabadi et al., 2020; Clark et al., 2019), debiased focal loss (Karimi Mahabadi et al., 2020), and other possible objectives, see Appendix B for more details. In experiments, we show that our framework can further improve these well-performing baseline models. One major difference with existing debiasing objectives is that prior works use a biased model to measure how much biases present in input, while FAIRFLOW encourages robust models to be undecided given known biased inputs, obtained by the proposed perturbations. Moreover, we do not impose any restriction on the parametrization of the underlying model f, making our framework flexible to work with a wide range of training methods and network architectures (Table 6-7 in Appendix).

3 Experiments

Setup We employ BERT (Devlin et al., 2019) as the commonly-used base model in previous works. In addition, we extend our evaluation to RoBERTa (Liu et al., 2019) and GPT-2 (Radford et al., 2019) for a more comprehensive analysis.

Datasets We evaluate our debiasing framework on three NLP datasets including MNLI (Williams et al., 2018), paraphrase identification using Quora question pairs (QQP) (Sharma et al., 2019), and relation extraction using gene-phenotype relation (PGR) (Sousa et al., 2019). These datasets are used for *in-domain* (ID) evaluation.

Stress Test Sets We assess the robustness of models against spurious correlations using "stress test sets," specifically designed with hard examples to challenge models. We use the stress test set for MNLI from (Naik et al., 2018), and use the same approach to generate the stress test set for QQP. For PGR, the label-preserving rules from previous tasks do not apply due to the nature of this dataset. However, given the long-tail distribution of entity appearances, we create a stress test set for PGR by selecting test examples in which both entities appear less than five times in the training set.

OOD Test Sets We assess the performance of models on existing out-of-distribution (OOD) test sets, which serve as another challenge benchmark. For MNLI, we use HANS (McCoy et al., 2019), which is designed to test models' capabilities against lexical and syntactic heuristics in data. For QQP, we employ the PAWS dataset (Zhang et al., 2019), which focuses on paraphrase identification in cases of high lexical and surface-level similarity between question pairs.

Transfer Test Sets We evaluate the performance of models in maintaining strong transferability across datasets. We use SNLI (Bowman et al., 2015) and MRPC (Dolan and Brockett, 2005) as the transfer set for MNLI and QQP, respectively.

Baselines We consider the following baselines:

- **FINETUNE** standard finetuning without debiasing based on the base model used.
- **E2E-POE** (Karimi Mahabadi et al., 2020), which trains a biased model on the hypothesis only and trains a robust model using Product of Experts (PoE) (Hinton, 2002).
- **DEBIASMASK** (Meissner et al., 2022), which first trains a weak learner and then prunes the robust model using PoE.
- **KERNELWHITENING** (Gao et al., 2022), which learns isotropic sentence embeddings using Nyström kernel approximation (Xu et al., 2015) method, achieving disentangled correlation between robust and spurious embeddings.
- **LWBC** (Kim et al., 2022), which learns a debiased model from a commitee of biased model obtained from subsets of data.
- **IEGDB** (Du et al., 2023), which mitigates dataset biases with an ensemble of random biased induction forest; the model induces a set of biased features and then purifies the biased features using information entropy[4].
- **READ** (Wang et al., 2023), which assumes that spuriousness comes from the attention and proposes to do deep ensemble of main and biased model at the attention level to learn robust feature interaction.

[4]While this method does not have a publicly released code, we tried our best to reproduce their approach and results with a few points lower than reported.

Model	MNLI (Acc.) ID	MNLI Stress	MNLI OOD	MNLI Transfer	QQP (F1) ID	QQP Stress	QQP OOD	QQP Transfer	PGR (F1) ID	PGR Stress	Avg. ID	Avg. Stress	Avg. OOD	Avg. Transfer
FINETUNE	84.3	61.7	59.7	78.7	88.6	63.3	47.7	65.1	64.3	55.2	79.1	60.1	53.7	71.9
DEBIASMASK	83.5	59.7	59.7	78.3	88.1	64.6	50.3	68.5	64.1	51.7	78.6	58.7	55.0	73.4
KERNELWHITENING	84.0	60.9	60.2	78.4	88.8	65.1	51.2	69.6	64.3	51.8	79.0	59.3	55.7	74.0
E2E-POE	83.4	61.3	62.3	77.5	88.5	64.5	51.4	70.5	63.0	53.6	78.3	59.8	56.8	74.0
LSWC	80.7	59.4	59.3	77.7	87.1	65.8	49.6	70.0	63.3	52.8	77.0	59.3	54.5	73.8
IEGDB	84.1	61.8	62.7	78.1	87.6	63.5	53.0	68.3	64.2	54.9	78.6	60.1	57.9	73.2
READ	80.8	61.5	63.4	75.1	87.0	66.7	53.6	68.2	63.0	54.4	76.9	60.9	58.5	71.7
FAIRFLOW-POE	84.6	64.3	64.3	79.5	88.8	71.0	53.9	70.4	64.9	55.9	79.4	63.7	59.1	<u>75.0</u>
FAIRFLOW-FOCAL	<u>84.9</u>	<u>64.8</u>	<u>64.3</u>	<u>79.3</u>	<u>89.5</u>	<u>71.3</u>	<u>54.9</u>	<u>70.7</u>	<u>65.4</u>	<u>56.5</u>	<u>79.9</u>	<u>64.2</u>	<u>59.6</u>	<u>75.0</u>
FAIRFLOW	**85.1**	**65.4**	**64.9**	**79.6**	**90.4**	**72.0**	**56.0**	**72.4**	**65.9**	**56.6**	**80.5**	**64.7**	**60.5**	**76.0**

Table 2: Experimental results on three datasets averaged across three architectures. Results for each architecture are shown in Table 5-7 in Appendix. The best performance is in **bold** and the second best is <u>underlined</u>.

4 Results and Discussions

Robust Debiasing Model The main results in **Table 2** shows our model with three objectives: contrastive learning (FAIRFLOW), product of experts (FAIRFLOW-POE) and focal loss (FAIRFLOW-FOCAL), see §2.4. They all achieve high performance across all datasets and test sets including in-domain (ID), stress, and out-of-distribution (OOD) test sets. By adopting the undecided learning objective, the model learns debiased and robust representations without loss of in-domain performance. Across three datasets, our best-performing model (FAIRFLOW) outperforms DEBIASMASK, KERNELWHITENING, E2E-POE, IEGDB, READ approaches by 2.0, 6.1 and 5.5; 1.5, 5.5 and 4.7; 2.2, 4.9 and 3.6; 1.9, 4.7 and 2.7; 3.6, 3.8 and 1.9 absolute points on the ID, stress and OOD test sets respectively. We attribute these gains to the use of biased branches and undecided learning, realized through the proposed contrastive objective.

We note that IEGDB and READ provide debiasing gains at the cost of ID performance, with a performance drop of 0.2, 1.0 and 0.1; 3.5, 0.4 and 1.3 compared to FINETUNE on MNLI, QQP, PGR respectively. Specifically, we attribute the large performance drop of READ to the deep ensemble (compared to logit ensemble of E2E-POE and FAIRFLOW-POE) of the target and biased model at the attention level, which may impose excessive regularization on the model. However, our model learns robust representations without loss on ID test sets across all three objectives.

In addition to better debiasing performance, our approach shows stronger transferability compared to baselines. Specifically, FAIRFLOW outperforms DEBIASMASK, KERNELWHITENING, and E2E-POE on transfer test set by 2.7, 2.1 and 2.1, respectively. In addition, FAIRFLOW-POE and FAIRFLOW-FOCAL retain strong transfer performance as well, indicating that our framework does not hurt models' transferability.

Comparing different fusion techniques in the last three rows in Table 2, we observe that the proposed contrastive objective is more effective than PoE (Karimi Mahabadi et al., 2020; Clark et al., 2019; Sanh et al., 2021) and debias focal loss (Karimi Mahabadi et al., 2020), in particular on stress and OOD test sets. We also find that debias focal loss almost always outperform PoE on our datasets, which is inline with previous report by Karimi Mahabadi et al. (2020).

More Bias Branches, Less Biased Model Unlike existing approaches that have a single view to dataset biases, our model employs multiple views, allowing it to effectively capture and mitigate various types of biases present in the data. Specifically, compared to E2E-POE which only captures one sub-input bias, FAIRFLOW-POE achieves on average 1.8, 9.5 and 3.8 absolute points improvement on ID, stress and OOD test set across three different datasets. Both methods employ PoE as the fusion technique. Compared to DEBIASMASK (Meissner et al., 2022) which only captures bias though a weak model, FAIRFLOW-POE achieves 1.5, 12.3 and 11.0 points improvement on ID, stress and OOD test sets, respectively.

Branches Contribute Differently To examine the contribution of each perturbation branch, we conduct ablation studies on MNLI. Specifically, we add one branch at a time to the vanilla model or remove one branch at a time from the full model, see **Table 3**. The perturbations include DropPremise and DropHypothesis, which drop the premise and hypothesis from the input respectively; HalfHalf, which randomly drops $k = 50\%$ of the tokens from input; Shuffle, which randomly shuffles the input; DropLayer, which drops all layers after the 2nd layer; and DestroyRep, which zeros out $m = 90\%$

Model	ID	Stress	OOD	Transfer
No debiasing	84.6	57.3	56.2	80.3
+ DropPremise	84.6	61.6	65.5	80.6
+ DropHypothesis	84.6	61.6	66.3	80.6
+ HalfHalf	84.8	62.1	64.2	80.0
+ Shuffle	84.8	62.1	63.9	80.0
+ DropLayer	84.8	62.0	65.4	80.4
+ DestroyRep	84.8	62.3	66.5	80.0
Full model	84.9	63.6	68.4	81.1
- DropPremise	84.6	61.6	63.2	80.6
- DropHypothesis	84.6	61.6	62.6	80.6
- HalfHalf	84.8	62.1	63.8	80.0
- Shuffle	84.8	62.1	65.3	80.0
- DropLayer	84.5	60.5	62.5	80.4
- DestroyRep	84.5	60.5	62.7	80.4

Table 3: Contribution of each perturbation branch in our method on MNLI.

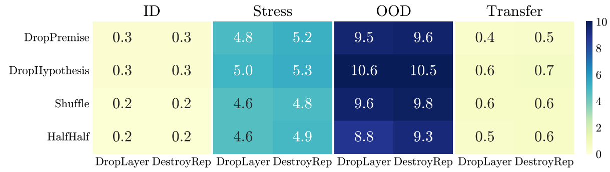

Figure 3: Debiasing performance with different combinations of explicit and implicit perturbations. The values indicate relative accuracy increase compared to vanilla fine-tuning.

Model	Param	Time (hr)
FINETUNE	110M + 2K	4.2
DEBIASMASK	+ 28M + 2K	5.3
KERNELWHITENING	+ 3K	6.3
E2E-POE	+ 30K	5.5
IEGDB	+ 50 × 2K	7.2
READ	+ 28M + 2K	4.9
FAIRFLOW	+ 2 × 2K	4.9

Table 4: Efficiency of debiasing models on MNLI.

of the elements in the intact representation. The results show that all perturbations contribute positively to the overall performance on ID, stress, OOD, and transfer test sets. Specifically, explicit perturbations can improve the vanilla model on average by 0.1 and 4.6 absolute point on ID and stress test sets respectively. While implicit perturbations improve the vanilla model on average by 0.1 and 4.9 points. In addition, DestroyRep achieves the best performance on the stress and OOD test sets, while DropPremise and DropHypothesis achieve the best performance on the transfer set.

In addition, we investigate the effect of different combinations of perturbations. Specifically, we train our model with one explicit perturbation and one implicit perturbation at a time. **Figure 3** illustrates the relative increase of performance to standard fine-tuning across ID, stress and OOD test sets. Two combinations yields better results on the OOD test set. The first combines DropPremise or DropHypothesis with DropLayer, while the second combines perturbation of all inputs (e.g. Shuffle) and PurturbRep. The improved results likely stem from the complementary strengths of these diverse perturbation techniques, which can create a more robust debiasing model.

Debiased Models Are Still Biased Our results in Table 2 and prior reports (Mendelson and Belinkov, 2021; Ravichander et al., 2023) show that debiased methods can still be biased. For example, DEBIASMASK and KERNELWHITENING show higher levels of biases than FINETUNE by 3.7 and 4.2 points on the stress test set (Naik et al., 2018) respectively. These results emphasize the need for modeling multiple types of biases, and highlights the advantages of our approach.

FAIRFLOW Maintains Generalization across Biases Bias in existing methods may be because of their tendency to over-specialize in specific types of biases. **Table 8** summarizes the performance of debiasing models across different subsets of the stress set. FAIRFLOW achieves the maximum average performance with smaller standard deviation across these subsets, indicating that it does not overfit to specific biases. We attribute such resilience to FAIRFLOW's incorporation of both explicit and implicit perturbations, along with the randomness in implicit perturbations, which allows the model to effectively handle diverse set of biases.

Efficiency We evaluate the efficiency of different debiasing methods in terms of number of trainable parameters and training time. As **Table 4** shows, FAIRFLOW introduces only 4K additional parameters, which is significantly less than 100K in IEGDB with 50 classifiers, and 28M in DEBIASMASK and READ with an extra weak model. This highlights the efficiency gains from the proposed perturbation operations. Furthermore, FAIRFLOW has the shortest training time. FAIRFLOW achieves these efficiencies without requiring additional training data, operating only by generating diverse views of the input data.

Perturbation for Data Augmentation The explicit perturbation operators proposed in our framework offer a valuable opportunity for data augmentation, leading to improved performances on existing debiasing methods (See Table 9 in Appendix).

Bias in Different Parts of Inputs In our experiments with single explicit perturbations, we find that DropPremise and DropHypothesis lead to similar performances on MNLI, showing that there exists dataset bias in premise, potentially as much as those in hypothesis. However, many existing methods tend to overlook biases in the premise in NLI datasets. In addition, biases can often emerge from the interplay of various parts of inputs, rather than a single source. HalfHalf and Shuffle perturbations can capture such types of biases by perturbing the entire inputs. We note that while additional weak learners can potentially capture biases from multiple sources (Utama et al., 2020b; Sanh et al., 2021; Meissner et al., 2022), their effectiveness is likely limited by the capabilities of the weak models. Our approach addresses dataset biases through a multi-view approach to bias, which leads to a more robust debiasing process.

5 Related Work

Quantifying Bias Several works focus on understanding dataset bias and deibasing algorithms, including measurement of bias of specific words with statistical test (Gardner et al., 2021), identification of biased and generation of non-biased samples with z-filtering (Wu et al., 2022), identification of bias-encoding parameters (Yu et al., 2023), when bias mitigation makes model less or more biased (Ravichander et al., 2023), bias transfer from other models (Jin et al., 2021), and representation fairness (Shen et al., 2022).

Debiasing with Biased Models These approaches model shortcuts from datasets, and use biased predictions as a reference to quantify bias in input data. Bias can be *explicit bias* in NLI datasets (Belinkov et al., 2019; Clark et al., 2019; Karimi Mahabadi et al., 2020; Utama et al., 2020a), and *implicit bias* detected by weak models (Ghaddar et al., 2021; Sanh et al., 2021; Meissner et al., 2022; Utama et al., 2020b; Meissner et al., 2022). Ensemble techniques include Product-of-Experts (PoE) (Hinton, 2002; Sanh et al., 2021; Cheng et al., 2024) which takes element-wise multiplication of the logits, Debiased Focal Loss (Karimi Mahabadi et al., 2020) and ConfReg (Utama et al., 2020a) which both down-weight predictions based on the confidence of biased models.

Debiased Representations Existing methods focus on weak-learner guided pruning (Meissner et al., 2022), disentangling robust and spurious representations (Gao et al., 2022), decision boundaries (Lyu et al., 2022), and attention patterns with PoE (Wang et al., 2023), training biased models with one-vs-rest approach (Jeon et al., 2023), and amplifying bias in training set with debiased test set (Reif and Schwartz, 2023).

Fairness and Toxicity These approaches focus on protected variable such as race. Existing methods spans across counterfactual data augmentation (Zmigrod et al., 2019; Dinan et al., 2020; Barikeri et al., 2021), comparisons between network architectures (Meade et al., 2022), deibasing with counterfactual inference (Qian et al., 2021), adversarial training (Madanagopal and Caverlee, 2023), prompt perturbation (Guo et al., 2023), data balancing (Han et al., 2022), contrastive learning (Cheng et al., 2021), detecting toxic outputs (Schick et al., 2021), performance degradation incurred by debiasing methods (Meade et al., 2022), and benchmarks (Nadeem et al., 2021; Hartvigsen et al., 2022; Sun et al., 2022). Social debiasing methods may underperform in OOD settings because OOD examples may not contain social stereotypes or biases.

6 Conclusion

We investigate bias mitigation in NLU datasets by formulating the debiasing problem within a contrastive learning framework, incorporating explicit and implicit perturbation techniques and introducing undecided learning. Through extensive experiments across a range of NLU tasks, we demonstrate the effectiveness of our method in achieving improved debiasing performance, while maintaining performance on in-domain test sets. We find that existing methods (including ours) are still sensitive to dataset biases, and our experiments show the limitations of these approaches in fully addressing dataset biases. These results necessitate investigating a more systematic evaluation benchmark for debiasing. Our approach can potentially be improved by investigating more complex biases (Yao et al., 2023; Gandikota et al., 2023), exploring alternative training paradigms such as curriculum learning (Bengio et al., 2009; Vakil and Amiri, 2022), and evaluating robustness to unseen biases (Tsirigotis et al., 2023). Beyond NLU, our work can potentially be applied to a broader range of applications (Cheng and Amiri, 2024; Liu et al., 2024).

Limitations

Though our framework outperforms baselines, there is still room for improvement on Stress and OOD test sets. In addition, we didn't analyze the generalizability of the approach to other NLP domains or tasks beyond the three tasks used in the experiments.

Ethic and Broader Impact Statements

Our research focuses on mitigating dataset biases in NLP datasets. There are no specific ethical concerns directly associated with this work. However, we acknowledge and emphasize the ethical mindfulness throughout the design, training, and applying the models investigated in this study on any applications. The broader impacts of our work are in advancing dataset fairness and potentially enhancing decision-making based on data. By addressing biases, we contribute to improving the reliability of NLP datasets and the accuracy and transferability of the models trained using NLP datasets.

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jiali Cheng and Hadi Amiri. 2024. Mu-bench: A multitask multimodal benchmark for machine unlearning. *arXiv preprint arXiv:2406.14796*.

Jiali Cheng, Mohamed Elgaar, Nidhi Vakil, and Hadi Amiri. 2024. Cognivoice: Multimodal and multilingual fusion networks for mild cognitive impairment assessment from spontaneous speech. In *Interspeech 2024*, pages 4308–4312.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Li Du, Xiao Ding, Zhouhao Sun, Ting Liu, Bing Qin, and Jingshuo Liu. 2023. Towards stable natural language understanding via information entropy guided debiasing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2868–2882, Toronto, Canada. Association for Computational Linguistics.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*.

SongYang Gao, Shihan Dou, Qi Zhang, and Xuanjing Huang. 2022. Kernel-whitening: Overcome dataset bias with isotropic sentence embedding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4112–4122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. End-to-end self-debiasing framework for robust NLU training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929, Online. Association for Computational Linguistics.

Qi Guo, Yuanhang Tang, Yawen Ouyang, Zhen Wu, and Xinyu Dai. 2023. Debias NLU datasets via training-free perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10886–10901, Singapore. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

Eojin Jeon, Mingyu Lee, Juhyeong Park, Yeachan Kim, Wing-Lam Mok, and SangKeun Lee. 2023. Improving bias mitigation through bias experts in natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11053–11066, Singapore. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. 2022. Learning debiased classifier with biased committee. In *Advances in Neural Information Processing Systems*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Xiulong Liu, Zhikang Dong, and Peng Zhang. 2024. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4478–4487.

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

Yougang Lyu, Piji Li, Yechang Yang, Maarten de Rijke, Pengjie Ren, Yukun Zhao, Dawei Yin, and Zhaochun Ren. 2022. Feature-level debiased natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Karthic Madanagopal and James Caverlee. 2023. Bias neutralization in non-parallel texts: A cyclic approach with auxiliary guidance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14265–14278, Singapore. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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

Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in NLU. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7607–7613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Ali Modarressi, Hossein Amirkhani, and Mohammad Taher Pilehvar. 2023. Guide the learner: Controlling product of experts debiasing method based on token attribution similarities. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1954–1959, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Jeff Pool and Chong Yu. 2021. Channel permutations for N: M sparsity. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13316–13327.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. 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 5434–5445, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abhilasha Ravichander, Joe Stacey, and Marek Rei. 2023. When and why does bias mitigation work? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9233–9247, Singapore. Association for Computational Linguistics.

Yuval Reif and Roy Schwartz. 2023. Fighting bias with bias: Promoting model robustness by amplifying dataset biases. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13169–13189, Toronto, Canada. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *ArXiv preprint*, abs/1907.01041.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. 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 7329–7346, Online. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1849–1857.

Diana Sousa, Andre Lamurias, and Francisco M. Couto. 2019. A silver standard corpus of human phenotype-gene relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1487–1492, Minneapolis, Minnesota. Association for Computational Linguistics.

Simeng Sun, Katherine Thai, and Mohit Iyyer. 2022. ChapterBreak: A challenge dataset for long-range language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3704–3714, Seattle, United States. Association for Computational Linguistics.

Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. 2023. Group robust classification without any group information. In *Advances in Neural Information Processing Systems*, volume 36, pages 56553–56575. Curran Associates, Inc.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Nidhi Vakil and Hadi Amiri. 2022. Generic and trend-aware curriculum learning for relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2202–2213, Seattle, United States. Association for Computational Linguistics.

Fei Wang, James Y. Huang, Tianyi Yan, Wenxuan Zhou, and Muhao Chen. 2023. Robust natural language understanding with residual attention debiasing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 504–519, Toronto, Canada. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Zenglin Xu, Rong Jin, Bin Shen, and Shenghuo Zhu. 2015. Nystrom approximation for sparse kernel methods: Theoretical analysis and empirical evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Implementation Details

For all datasets, we train all methods on the BERT-base (Devlin et al., 2019) checkpoint, with a 2e-5 learning rate with linear decay using AdamW (Kingma and Ba, 2015) optimizer. The batch size is set to 32. For the baseline models, we follow their papers for the hyperparameter choices. All experiments on done on a single A100 GPU.

We implement the proposed perturbation as illustrated in Table 1 by randomly dropping 50% of the tokens from each sentence, dropping all layers after the second layer (3–12), and zeroing $m = 90\%$ of the elements in the intact representation $f(x_i)$.

Each branch-specific MLP consists of two linear layers with a ReLU activation function in between. We use $\lambda = 0.1$ in our experiments.

B Other Debiasing Objectives

The idea of existing debiasing objectives is based on the idea of adjusting the importance of training examples, i.e. their contribution to loss calculation. The importance of examples which the model fails the correctly predict is promoted while the importance of examples which the model correctly predicts is reduced.

Product-of-Experts (PoE) (Clark et al., 2019; Karimi Mahabadi et al., 2020; Sanh et al., 2021) is one of the most commonly adopted debiasing objective, which takes dot product of the logits of the main model and the biased models. Debiasing Focal Loss (Karimi Mahabadi et al., 2020) downweights the main model based on how close the logits of the biased models is to 1. Confidence Regularization (Utama et al., 2020b) reduced the loss scale of examples with a scaling mechanism.

C RoBERTa as Encoder

We conducted experiments on RoBERTa-base (Liu et al., 2019) using the MNLI dataset to evaluate the efficacy of FairFlow more effectively. The results in Table 6 shows that the performance of all models improved using RoBERTa-base as encoder. We also observe comparable gains to BERT as encoder in case of ID and Transfer settings and smaller gains in case of Stress and OOD settings, which can be attributed to the use of a more powerful encoder.

D Perturbation for Data Augmentation

The explicit perturbation operators proposed in our framework offer a valuable opportunity for data augmentation. This can be particularly useful in tasks such as NLI. Consider the example (x_i^p, x_i^h, y_i), where x_i^p represents the premise, x_i^h represents the hypothesis, and y_i denotes the label. To augment the dataset, we create additional data samples by applying different perturbation operations, e.g., by dropping the premise: (‘’, x_i^h, not entailment), dropping the hypothesis: (x_i^p, ‘’, not entailment), shuffling the data: ($\mathcal{P}_{Irr}(x_i^p)$, $\mathcal{P}_{Irr}(x_i^h)$, not entailment) and dropping parts of the input: ($\mathcal{P}_{Sub}(x_i^p)$, $\mathcal{P}_{Irr}(x_i^h)$, not entailment). The augmented examples can be added back to the original dataset to mitigate the effect of bias during fine-tuning and potentially enhance model's generalizability, leading to improved performance on existing debiasing methods (See Table 9).

Model	MNLI (Acc.)				QQP (F1)				PGR (F1)	
	ID	Stress	OOD	Transfer	ID	Stress	OOD	Transfer	ID	Stress
FINETUNE	84.4	55.8	60.7	80.1	89.1	59.3	40.8	61.8	67.1	54.3
DEBIASMASK	84.7	53.6	60.8	80.5	88.3	60.2	44.7	62.1	65.4	44.6
KERNELWHITENING	83.3	53.5	60.5	80.2	87.6	61.3	45.1	62.7	63.5	42.0
E2E-POE	83.8	57.8	66.3	80.1	89.2	58.9	42.5	63.1	63.2	50.3
LWBC	83.2	58.3	60.2	80.7	89.6	73.2	49.2	67.4	66.5	53.2
IEGDB	84.5	60.1	67.2	79.8	84.6	57.3	50.6	60.5	64.8	54.6
READ	79.6	58.3	**68.4**	73.0	84.5	65.8	46.7	61.7	62.6	55.0
FAIRFLOW-POE	84.8	62.3	67.5	81.0	89.2	77.5	48.9	63.1	67.4	55.6
FAIRFLOW-FOCAL	84.8	62.8	67.9	80.9	89.6	77.8	49.2	63.1	**67.7**	**56.1**
FAIRFLOW	**84.9**	**63.6**	**68.4**	**81.1**	**91.8**	**78.4**	**51.5**	**68.3**	**67.7**	55.8

Table 5: Experimental results on three datasets using BERT as the base model. The best performance is in **bold** and the second best is underlined. Note that IEGDB does not release their code. We tried our best to reproduce the results but failed on HANS, which is 5.2 points lower than the reported 72.4. This is potentially due to implementation and optimization details which the authors did not release.

Model	MNLI (Acc.)				QQP (F1)				PGR (F1)	
	ID	Stress	OOD	Transfer	ID	Stress	OOD	Transfer	ID	Stress
FINETUNE	88.1	75.3	66.4	81.0	92.2	63.5	44.7	68.3	69.3	57.1
DEBIASMASK	86.5	72.7	66.9	80.7	92.5	66.1	49.1	68.7	70.2	57.5
KERNELWHITENING	88.1	74.1	67.4	79.9	93.1	66.7	50.2	68.9	71.3	58.3
E2E-POE	88.3	72.6	69.5	80.7	92.4	66.4	50.3	68.5	70.5	57.9
LWBC	84.6	69.3	66.7	81.0	91.7	63.2	43.9	67.4	70.4	54.2
IEGDB	88.2	72.4	69.3	80.3	92.3	66.3	50.2	68.3	70.8	56.3
READ	85.3	73.5	70.3	78.5	91.4	68.1	51.0	67.8	69.3	55.7
FAIRFLOW-POE	88.3	76.1	70.2	81.4	92.5	66.7	50.6	68.3	70.8	58.0
FAIRFLOW-FOCAL	88.2	76.7	70.3	81.4	92.7	67.8	51.3	68.7	71.1	58.3
FAIRFLOW	88.3	77.2	70.4	81.2	93.3	68.4	51.8	68.6	71.4	58.3

Table 6: Results using RoBERTa (Liu et al., 2019) as the base model. The best performance is in **bold** and the second best is underlined.

Model	MNLI (Acc.)				QQP (F1)				PGR (F1)	
	ID	Stress	OOD	Transfer	ID	Stress	OOD	Transfer	ID	Stress
FINETUNE	80.4	54.0	52.1	75.2	84.5	67.3	57.7	65.1	56.5	54.2
DEBIASMASK	79.3	52.8	51.6	73.7	83.6	67.5	57.3	74.9	56.7	53.2
KERNELWHITENING	80.7	55.1	52.8	75.1	85.7	67.4	58.3	77.2	58.3	55.1
E2E-POE	78.1	53.6	51.1	71.9	83.9	68.2	61.5	80.1	55.4	52.7
LWBC	74.5	50.8	51.0	71.4	80.2	61.0	55.8	75.4	53.2	51.0
IEGDB	79.7	52.9	51.8	74.3	86.1	66.9	58.2	76.2	57.1	53.8
READ	77.5	52.7	51.5	73.8	85.2	66.4	63.3	75.2	57.3	52.6
FAIRFLOW-POE	80.9	54.7	55.2	76.2	84.7	68.8	62.3	80.0	56.5	54.1
FAIRFLOW-FOCAL	81.8	55.1	54.9	75.8	86.3	68.5	64.2	80.4	57.4	55.3
FAIRFLOW	82.2	55.6	56.1	76.7	86.3	69.3	64.7	80.4	58.6	55.8

Table 7: Results using GPT-2 as the base model. The best performance is in **bold** and the second best is underlined.

Model	Avg. Acc (↑)	Std. Acc (↓)
FINETUNE	60.1	9.3
DEBIASMASK	58.7	6.7
KERNELWHITENING	59.3	5.9
E2E-POE	60.0	6.1
LSWC	59.4	5.8
IEGDB	60.1	7.3
READ	60.9	5.6
FAIRFLOW-POE	63.8	5.7
FAIRFLOW-FOCAL	64.3	5.2
FAIRFLOW	**64.7**	**5.1**

Table 8: Average performance and standard deviation on each type of stress test averaged across three architectures. The best performance is in **bold** and the second best is underlined.

Model	MNLI (Acc.)			
	ID	Stress	OOD	Transfer
FINETUNE	84.4	55.8	60.7	80.1
FINETUNE + Aug	84.5	59.1	61.0	81.0
DEBIASMASK	84.7	53.6	60.8	80.5
DEBIASMASK + Aug	85.6	55.4	62.2	81.1
KERNELWHITENING	83.3	53.5	60.5	80.2
KERNELWHITENING + Aug	85.1	56.2	60.8	81.0
E2E-POE	83.8	57.8	66.3	80.1
E2E-POE + Aug	84.8	61.1	66.2	80.6
IEGDB	84.5	60.1	65.7	79.8
IEGDB + Aug	85.6	60.8	66.4	80.7
READ	79.6	58.3	68.4	73.0
READ + Aug	79.6	58.3	69.6	77.2

Table 9: Performance when applying data augmentation, which effectively improve existing debiasing methods. The best performance is in **bold**.