Deep Learning on a Healthy Data Diet: Finding Important Examples for Fairness

Abdelrahman Zayed[1,2*], **Prasanna Parthasarathi**[1,3], **Gonçalo Mordido**[1,2], **Hamid Palangi**[4], **Samira Shabanian**[4†], **Sarath Chandar**[1,2,5†]

[1]Mila - Quebec AI Institute
[2]Polytechnique Montreal
[3]McGill University
[4]Microsoft Research
[5]Canada CIFAR AI Chair
{zayedabd,parthapr,goncalo-filipe.torcato-mordido,sarath.chandar}@mila.quebec
{hpalangi,samira.shabanian}@microsoft.com

Abstract

Data-driven predictive solutions predominant in commercial applications tend to suffer from biases and stereotypes, which raises equity concerns. Prediction models may discover, use, or amplify spurious correlations based on gender or other protected personal characteristics, thus discriminating against marginalized groups. Mitigating gender bias has become an important research focus in natural language processing (NLP) and is an area where annotated corpora are available. Data augmentation reduces gender bias by adding counterfactual examples to the training dataset. In this work, we show that some of the examples in the augmented dataset can be not important or even harmful to fairness. We hence propose a general method for pruning both the factual and counterfactual examples to maximize the model's fairness as measured by the demographic parity, equality of opportunity, and equality of odds. The fairness achieved by our method surpasses that of data augmentation on three text classification datasets, using no more than half of the examples in the augmented dataset. Our experiments are conducted using models of varying sizes and pre-training settings. *WARNING: This work uses language that is offensive in nature.*

1 Introduction

Although pre-trained language models (Vaswani et al. 2017; Devlin et al. 2019; Radford et al. 2018, 2019; Brown et al. 2020; Raffel et al. 2020; Yang et al. 2019) have been tested on a variety of language understanding and generation benchmarks (Wang et al. 2018, 2019; Rajpurkar et al. 2016; Budzianowski et al. 2018; Zhang et al. 2018; Mordido and Meinel 2020; Sauder et al. 2020), the fairness of these models with respect to marginalized communities has recently come under scrutiny (Dixon et al. 2018; Zhang et al. 2020; Garg et al. 2019). In terms of unintended gender bias in prediction, the works by Nadeem, Bethke, and Reddy (2021) and Meade, Poole-Dayan, and Reddy (2022) show that pre-trained language models, such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), GPT-2 (Radford et al. 2019), and XLNet (Yang et al. 2019) behave differently based on the presence or absence of gender cues in the input text, while the prediction should have remained agnostic.

Similar results have been shown for non-pretrained recurrent networks (Hall Maudslay et al. 2019; De-Arteaga et al. 2019; Lu et al. 2020) such as long short-term memory models (LSTMs) (Hochreiter and Schmidhuber 1997) and gated recurrent units (GRUs) (Chung et al. 2014). Kiritchenko and Mohammad (2018) also highlight that support vector machines (SVMs) (Hearst et al. 1998) applied for sentiment analysis predict *anger* with a higher probability for input texts having references to female gender (*e.g. she*, *her*, *woman*, *lady*) over texts with male references (*e.g. he*, *him*, *himself*, *man*), with the same context. A biased model should not replace humans in certain tasks (*e.g.* resume filtering or loan eligibility prediction), regardless of how accurate it is if it achieves high accuracy on a test set that is not representative of the population. For example, if most test examples refer to men, this could hide the model's poor performance on examples that reference women. Hence, relying solely on metrics such as accuracy and $F1$ might be misleading.

To mitigate gender bias in language models, several techniques have been proposed. These methods can be broadly classified into three main categories: data-based methods (Lu et al. 2020; Hall Maudslay et al. 2019; De-Arteaga et al. 2019); regularization-based methods (Gupta et al. 2021; Garg et al. 2019); and adversarial-based methods (Song et al. 2019; Madras et al. 2018; Jaiswal et al. 2020). Data-based methods, which are our focus in this work, change the training data to balance the bias through targeted data augmentation (Lu et al. 2020; Hall Maudslay et al. 2019), constructing counterfactual examples (Garg et al. 2019), or removing protected attributes from the input (De-Arteaga et al. 2019) to disallow models from learning any correlation between labels and gender words. Regularization-based methods add an auxiliary loss term to the objective function that reduces the amount of bias in the model. Adversarial-based methods use adversarial learning for bias mitigation.

Although data-based methods are effective in theory, they are also demanding in different ways. First, some methods such as counterfactual data augmentation (CDA) (Lu

[*]Work done during an internship at Microsoft Research.
[†]Equal advising.

et al. 2020) add counterfactual examples to the training data, which substantially increases the training time. Second, data balancing methods (Dixon et al. 2018) manually collect more examples for the under-represented groups, which requires human intervention. Third, the performance may degrade on the main downstream task (Zhang et al. 2020; Meade, Poole-Dayan, and Reddy 2022). In this work, we propose the gender equity (*GE*) score, which ranks the counterfactual examples based on their contribution to the overall fairness of the model. Intuitively, this score makes data augmentation methods more efficient by only using the examples that have the largest contribution to fairness. The pruning nature of our approach also helps exclude the harmful examples that degrade the overall fairness of the model by enforcing stereotypical associations about different genders. Our goal is to find the best trade-off between fairness and performance. Our contributions are summarized as follows:

1. We propose a way to filter the examples that enforce undesired gender stereotypes from the training data, thus improving the overall fairness of the model compared to conventional data augmentation.
2. We reduce the redundancy in counterfactual data augmentation by pruning the counterfactual examples that are not important for fairness (*i.e.* the ones that do not contain gender words).
3. We study the effect of gender bias mitigation on downstream task performance and find that our method only shows a degradation of no more than 3% on the AUC over three popular language understanding tasks that are concerned with biases and stereotypes, compared to the original (biased) BERT and RoBERTa models.

2 Related Work

Gender bias Gender bias is defined as the tendency of the system (which is the machine learning model in our case) to change its prediction based on the gender of the person referred to in the sentence (Friedman and Nissenbaum 1996). Existing works that study gender bias can be categorized into structural (Adi et al. 2017; Hupkes, Veldhoen, and Zuidema 2018; Conneau et al. 2018; Tenney, Das, and Pavlick 2019; Belinkov and Glass 2019) or behavioural (Sennrich 2017; Isabelle, Cherry, and Foster 2017; Naik et al. 2018) approaches. Our work follows a behavioral approach for bias quantification. Structural methods measure gender bias by focusing on the embeddings that the model assigns to sentences or words, regardless of the model's prediction (Tenney, Das, and Pavlick 2019; Belinkov and Glass 2019). One of the earliest structural methods to quantify bias is known as the **W**ord **E**mbedding **A**ssociation **T**est (WEAT) (Brunet et al. 2019), which measures the similarity between the word embeddings of some target words (such as "men" and "women") and some attribute words (such as "nice" or "strong"). According to this metric, a model is considered biased if the word embeddings for masculine words such as "man" and "boy" have a higher cosine similarity with attribute words such as "strong" and "powerful" than feminine words such as "woman" and "girl". The work by May et al. (2019) extended this concept to sentence embeddings by introducing a **S**entence **E**mbedding **A**ssociation **T**est (SEAT).

Behavioral methods, on the other hand, measure the model's bias based on its predictions on synthetic datasets designed specifically for gender bias assessment. The work by Dixon et al. (2018) measured bias using the **F**alse **P**ositive **E**quality **D**ifference (FPED) and **F**alse **N**egative **E**quality **D**ifference (FNED). In other words, they measure bias as the inconsistency in false positive/negative rates across different genders. For example, if a model has a higher false positive/negative rate when the input examples refer to women, then the model is biased.

Data-based bias mitigation methods One of the early works in data-based bias mitigation methods is the work by Dixon et al. (2018), which proposed an expensive, yet effective, data augmentation technique through manual labeling to account for identity balancing issues. Lu et al. (2020) proposed counterfactual data augmentation (CDA) which requires constructing a counterfactual example for every example in the dataset, thus doubling the size of the training data. Hall Maudslay et al. (2019) proposed counterfactual data substitution (CDS) where the model replaces each example with its corresponding counterfactual example based on a fair coin toss, which keeps the training data size constant. The authors proposed also to replace names, so they refer to another gender. De-Arteaga et al. (2019) proposed not exposing the model to gender tokens by simply removing them from the dataset.

Another similar approach is called counterfactual logit pairing (Garg et al. 2019), which involves creating a counterfactual sentence for every input sentence, such that the model is encouraged to give the same predictions for both. For any input sentence, the corresponding counterfactual sentence carries the same meaning while referring to an alternative identity group. For example, a sentence like "he is a gay man" may become "he is a straight man". The intuition here is to teach the model not to base its decision on identity characteristics such as sexual orientation, nationality, and gender. Although the authors empirically show the effectiveness of this method, it is limited by the automatic replacement of a group of 50 words representing different identity groups (Dixon et al. 2018). Among those 50 words, only the words "male" and "female" are related to gender, which restricts the method's applicability to sentences containing these exact words. For example, gender pronouns, which are strong indicators of gender, are ignored.

Performance-based data pruning methods Toneva et al. (2019) propose ranking the examples of a given training dataset based on the number of times they are forgotten during training, denoted as the *forget score*. Forgetting happens when the example is classified incorrectly after it had been correctly classified in a previous epoch. Intuitively, the examples that are forgotten more often during training are more important and should be kept, while those that are never forgotten could be removed from the dataset without affecting the model's performance. On the other hand, Paul, Ganguli, and Dziugaite (2021) propose to rank the training examples based on the ℓ_2-norm of the difference between

the prediction of the model and the ground truth, denoted as the *EL2N* score. The larger the norm of the error, the more important the example is. The same authors also propose to rank the training examples based on the mean of the ℓ_2-norm of the gradient of the loss function with respect to the weights, denoted as the *GraNd* score. Both scores have a strong correlation and are calculated early in training.

3 Background

We define a text classification task on a dataset D of N examples, such that $D = \{\mathrm{s}_0, \mathrm{s}_1, \ldots, \mathrm{s}_{N-1}\}$ where s_i is the i^{th} example. Each sentence s_i is a composition of m tokens, where m is the maximum number of tokens in any sentence, such that $\mathrm{s}_i = \left\{w_i^0, w_i^1, \ldots, w_i^{m-1}\right\}$, where w represents the token. The objective of the task is to learn a classifier, parameterized by θ, to output a label $y_i \in \{0, 1\}$, where y_i is 1 when s_i is a toxic/sexist sentence, and 0 otherwise. The optimal set of parameters θ^* for the model is found using maximum likelihood, as follows:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_{i=0}^{N-1} \mathcal{L}_\theta \left(y_i \mid \phi\left(\mathrm{s}_i\right)\right) \tag{1}$$

where $\phi\left(\mathrm{s}_i\right)$ yields the sentence embedding x_i from the sentence s_i, such that $\phi : \mathrm{s}_i \rightarrow \mathrm{x}_i \in \mathbb{R}^{K \times m}$, K is the dimension of the word embedding, and $\mathcal{L}_\theta$ is the likelihood of the predictor function parameterized by θ.

To understand the bias in the classifier, we disentangle the intent of a sentence s_i from the gender identity denoted by z_i. For example, if $\mathrm{s}_i|do(z = z_i)$ is "he explained the situation to her" then $\mathrm{s}_i|do(z = \neg z_i)$ would be "she explained the situation to him", where $\mathrm{s}_i|do(z = \neg z_i)$ represents the same sentence s_i had the gender words been flipped (Pearl 1995). This is referred to as the counterfactual sentence. Based on the previously mentioned notations, a model, parameterized by θ, is considered biased if its output y depends on z, *i.e.*, $y \not\perp\!\!\!\perp z$.

4 Deep Learning on a Healthy Data Diet

Building on the data diet approach introduced by Paul, Ganguli, and Dziugaite (2021), we propose a "healthier" version of the data diet that optimizes both fairness and performance using a reduced set of examples. The proposed *GE* score is inspired by the error norm score (*EL2N*) (Paul, Ganguli, and Dziugaite 2021), which determines the importance of an example for performance by how far its prediction is from the ground truth. Similarly, we determine the importance of an example for fairness by how far its prediction is from the prediction when the gender words in the input sentence are flipped. We do so by estimating the norm of the difference between the prediction for factual examples and the corresponding counterfactual examples. Our score is defined as:

$$\textit{GE}(\mathrm{s}_i) = ||f_\theta\left(\mathrm{s}_i|do(z = z_i)\right) - f_\theta\left(\mathrm{s}_i|do(z = \neg z_i)\right)||_2 \tag{2}$$

where $f_\theta\left(\mathrm{s}_i\right)$ represents the logit outputs of the model for the i-th input sentence s_i and $\|\cdot\|_2$ is the ℓ_2-norm. The $do(\cdot)$ indicates whether the logits are obtained using the actual gender indication ($z = z_i$) or its counterfactual version ($z = \neg z_i$). Note that examples that do not contain gender words will get a score of zero since we assume they do not contribute to the model fairness with respect to gender. The intuition behind the *GE* score is that if the model's prediction changes drastically upon changing the gender words in a sentence, then this reveals that the sentence contains gender words that the model correlates with the output label. In this case, we add the counterfactual example with the same ground truth label as the factual example to help mitigate this undesirable effect by teaching the model that the output should remain the same independently of the gender words.

It is important to mention that we flip names (for instance, *John* to *Alice*), gender pronouns, as well as gendered words such as *king* and *queen*. Based on this definition, our *GE* score for any sentence s_i is the same as the score for its counterfactual. Although gender is not binary (Manzini et al. 2019; Dinan et al. 2020), we assume it to be binary in this work for simplicity. We intend to extend our method to non-binary gender in future work.

4.1 Finding Important Counterfactual Examples for Fairness

The *GE* score is used to measure the importance of every counterfactual example in the dataset for fairness. With this information, we propose to train with only the most important examples for fairness to perform bias mitigation. Similar to the *EL2N* score, the *GE* score is computed during the early stages of training and averaged over multiple initializations. The number of epochs used to compute our score is a small fraction of the total number of epochs needed for convergence. It is important to note that *GE* score is both dataset and model-dependent. We study the effect of changing the dataset, model architecture and initialization, as well as the number of the early stage epochs after which the score is computed in section A.1 of the technical appendix.

4.2 Combining the Factual and Counterfactual Examples

CDA includes both the factual as well as the counterfactual examples in the training dataset, while CDS proposes replacing the factual version of every example with its counterfactual one with a probability 0.5. In this work, we provide a recipe for a novel way of combining the factual and counterfactual examples such that we outperform the fairness obtained by CDA and CDS, as measured by the demographic parity (DP), equality of opportunity for $y = 1$ (EqOpp1), and equality of odds (EqOdd).

We choose $a\%$ and $b\%$ from the factual and counterfactual examples, respectively, and prune the rest. The $a\%$ from the factual examples are chosen randomly, while the $b\%$ chosen from the counterfactual examples are the ones with the highest GE score, as shown in Fig. 1. The intuition is that the counterfactual examples with high GE scores are mostly the ones that mitigate the stereotypical biases that the model might have learned during pre-training (Vig et al. 2020; Nadeem, Bethke, and Reddy 2021), so they are useful for fairness. We repeat this for different combinations of $a\%$ and $b\%$ to find the one that achieves the highest fairness.

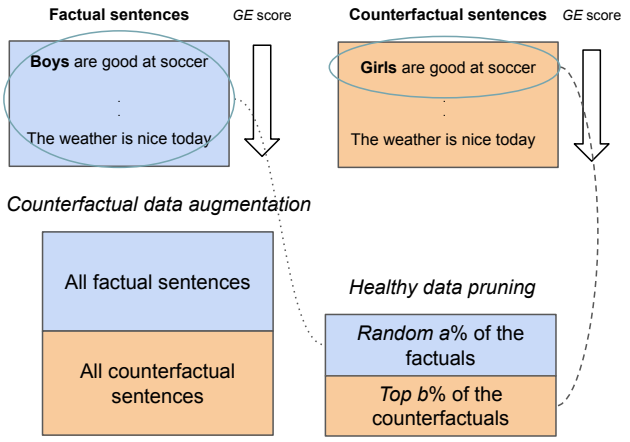

Figure 1: A comparison between our healthy pruning method and conventional CDA, which uses all the counterfactual and factual examples. Healthy data pruning reduces the number of examples needed to promote fairness. The counterfactual examples with high GE scores mitigate the stereotypical correlations that the model might have learned during pre-training, so we prioritize adding them.

5 Experiments

In this section, we describe the tasks, datasets, baselines, and evaluation metrics used in our experiments. Overall, we showcase how our proposed *GE* score reflects the importance of the counterfactual examples on multiple models and datasets. Moreover, we study how *GE* score may be leveraged to achieve a good trade-off between performance and fairness.

5.1 Datasets

We consider two different binary text classification tasks — sexism and toxicity detection. The tasks are to train a model to appropriately distinguish texts that are sexist/toxic from the ones that are not. Following Dixon et al. (2018), a toxic comment is defined as a comment that leads a person to leave a discussion. In all previously mentioned tasks, the model should base its predictions on the meaning of the sentence, rather than the gender of the person it refers to. We use the following three datasets:

1. Twitter dataset (Waseem and Hovy 2016): This dataset is composed of approximately 16K tweets, that are classified as racist, sexist, or neither racist nor sexist. We binarized the labels to only sexist and not sexist, by merging the racist tweets with the non-sexist tweets.
2. Wikipedia dataset (Dixon et al. 2018): This dataset is composed of approximately 160K comments labeled as toxic or not toxic, which are extracted from Wikipedia Talk Pages [1].
3. Jigsaw dataset: This dataset is composed of approximately 1.8M examples taken from a Kaggle competition[2]. The original task is to classify the input sentence to one of five different labels for different degrees of toxicity, but we binarized the labels to only toxic and not toxic by merging all non-toxic examples. We down-sampled the dataset to 125k examples to accommodate for the available computational resources.

An overview of each dataset is presented in Table 1. Specifically, we show the size of each dataset in terms of the number of sentences in the training data, the number of occurrences of gender pronouns, and the percentage of positive (*i.e.* sexist/toxic) examples. It is worth mentioning that gender words are detected using the *gender-bender* Python package [3]. An illustration of how our *GE* score measures the importance of several counterfactual examples in terms of fairness from the Twitter and Wikipedia datasets is presented in Table 2.

5.2 Baselines

For relative comparison of our proposed method, we use (1) a vanilla model that is not trained using any bias mitigation heuristics, (2) counterfactual data augmentation (CDA) (Lu et al. 2020), (3) counterfactual data substitution (CDS) (Hall Maudslay et al. 2019). We also tried data balancing (Dixon et al. 2018) and gender blindness (De-Arteaga et al. 2019), but we found them not effective for the specific datasets and models that we used. Similar to the work of Hall Maudslay et al. (2019), our implementation of CDA includes name flipping.

5.3 Evaluation Metrics

To measure fairness, we report the demographic parity (DP) (Beutel et al. 2017; Hardt, Price, and Srebro 2016), which is defined as:

$$DP = 1 - |p(\hat{y} = 1|z = 1) - p(\hat{y} = 1|z = 0)| \quad (3)$$

where $\hat{y}$ refers to the model's prediction and $z \in \{0, 1\}$ refers to keeping and flipping the gender words in the sentence, respectively. To compute the DP, we follow the procedure in other works (Dixon et al. 2018; Park, Shin, and Fung 2018) that use a synthetic dataset called the Identity Phrase Templates Test Set (IPTTS) for measuring fairness metrics. We also use EqOdd and EqOpp1 as two additional fairness metrics, which we define in section A.2 in the technical appendix. We use the area under the receiver operating characteristic curve (AUC) as our performance metric.

5.4 Experiment Details

We train the model for 15 epochs using cross-entropy loss for both the Twitter and Wikipedia datasets, and 10 epochs for the Jigsaw dataset. We used a ratio of 8:1:1 between training, validation, and test for all datasets, except for the Wikipedia toxicity dataset, where we followed the split ratio used by Dixon et al. (2018). For the *GE* score, we chose the early stages of training to be the state of the model after one

[1] https://www.figshare.com/articles/dataset/Wikipedia_Talk_Labels_Toxicity/4563973

[2] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[3] https://github.com/Garrett-R/gender_bender

Dataset	Size	Male pronouns	Female pronouns	Positives
Twitter	16,907	901	716	20.29%
Wikipedia	159,686	54,357	11,540	9.62%
Jigsaw	125,000	51,134	13,937	5.98%

Table 1: Statistics of the three datasets used in our experiments. The male pronouns considered are *he*, *him*, *himself*, and *his*, while the female pronouns are *she*, *her*, *hers*, and *herself*.

Factual	Counterfactual	*GE*
okay king of the Wikipedia Nazis	okay queen of the Wikipedia Nazis	0.36
Kate you stupid woman!	Kareem you stupid man!	0.11
I'm not sexist.. But women drivers are terrible	I'm not sexist.. But men drivers are terrible	0.10
Oh my god.... When will this show end	Oh my god.... When will this show end	0.00

Table 2: Examples of sentences from the Twitter and Wikipedia datasets with their *GE* score show how important it is to include their counterfactual examples in the training data for bias mitigation. The *GE* scores are obtained using BERT model.

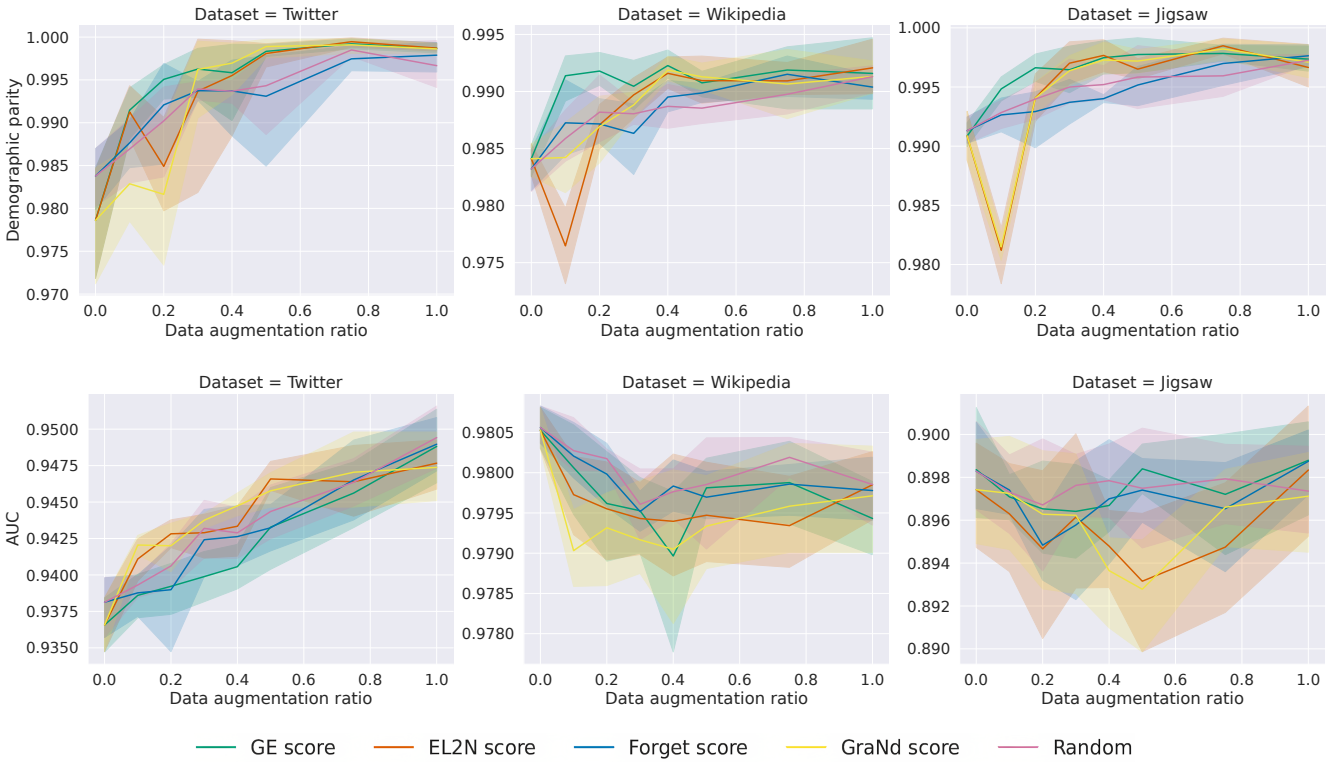

Figure 2: DP and AUC of BERT using different ratios and rankings of the counterfactual examples for data augmentation on the three datasets. The ratio represents the top b % of the counterfactual examples. All plots are best viewed in color.

epoch. More details about the effect of selecting the epoch for calculating the *GE* score are presented in section A.1 in the technical appendix. All the results are obtained by running the experiments for five different seeds. We used BERT and RoBERTa base models for text classification. Section A.3 in the technical appendix includes all the details needed about the hyper-parameters used to obtain the results. Our code is publicly available for reproducibility [4]. Sections B.1, B.2, and B.3 in the code appendix provide more details about the pre-processing of the datasets, the procedure to conduct and analyze the experiments, as well as the computing infrastructure used for running the experiments, respectively.

Experiment 1: Verifying that *GE* score reflects the importance of counterfactual examples In this experiment, we want to verify empirically that our *GE* score reflects the importance of each counterfactual example for fairness. Therefore, we train the model on all the factual data, as well as a varying ratio of the counterfactual examples. The counterfactual examples are ranked based on their *GE* scores,

[4] https://github.com/chandar-lab/healthy-data-diet

Figure 3: DP and AUC of RoBERTa using different ratios and rankings of the counterfactual examples for data augmentation on the three datasets. The ratio represents the top b % of the counterfactual examples.

Ranking for the factuals	Ranking for the counterfactuals	Ranking name
Ascending GE score	Ascending GE score	Vanilla GE
Random	Random	Random
Random	Descending *GE* score	Healthy random
Random	Ascending GE score	Unhealthy random

Table 3: The different ranking methods that we consider to prune the dataset in our experiments.

such that we only use the top $b\%$ and prune the rest, where $b \in \{0, 0.1, .., 1\}$. We compared our *GE* score with existing performance-based example scores, namely the (1) *EL2N* score, (2) *GraNd* score, (3) *forget score*, as well as the (4) random score. Figs. 2-3 show that ranking the counterfactual sentences based on the *GE* score results in a faster increase in the DP, compared to other ranking scores. However, using RoBERTa on the Wikipedia dataset, both the *EL2N* and *GraNd* scores have almost the same rate of increase in DP as the *GE* score. We hypothesize that this is due to a considerable overlap between the set of examples that are important for fairness (picked up by the *GE* score) and the set of examples important for performance (picked up by the *EL2N* and *GraNd* scores), for this specific model and dataset. We verify and discuss our hypothesis in more detail in section A.4 in the technical appendix. Since we observe that the improvement in DP achieved by ranking the counterfactual examples based on the *GE* score is often accompanied by a degradation in the AUC, we explore the trade-off between fairness and performance in the next experiment.

Experiment 2: Finding the best trade-off between fairness and performance We choose a and b % of the samples from the factual and counterfactual examples, respectively, with $a \in \{0.3, 0.4, 0.5\}$ and $b \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Table 3 shows the different settings that we consider to rank the factual and counterfactual examples. For every setting, Fig. 4 shows the best configuration among all the different combinations of $a\%$ and $b\%$, based on the highest improvement in DP such that the degradation in AUC over the biased model is not more than 3%. It is clear that the model's fairness is affected by the ranking of the factual and counterfactual examples. When the factual examples are chosen randomly and the counterfactual examples are ranked based on the ascending *GE* score, we get the worst fairness, which we refer to as the "unhealthy random" ranking. This occurs because we choose the least important counterfactual examples. The opposite happens when choosing the factual examples randomly and the counterfactual

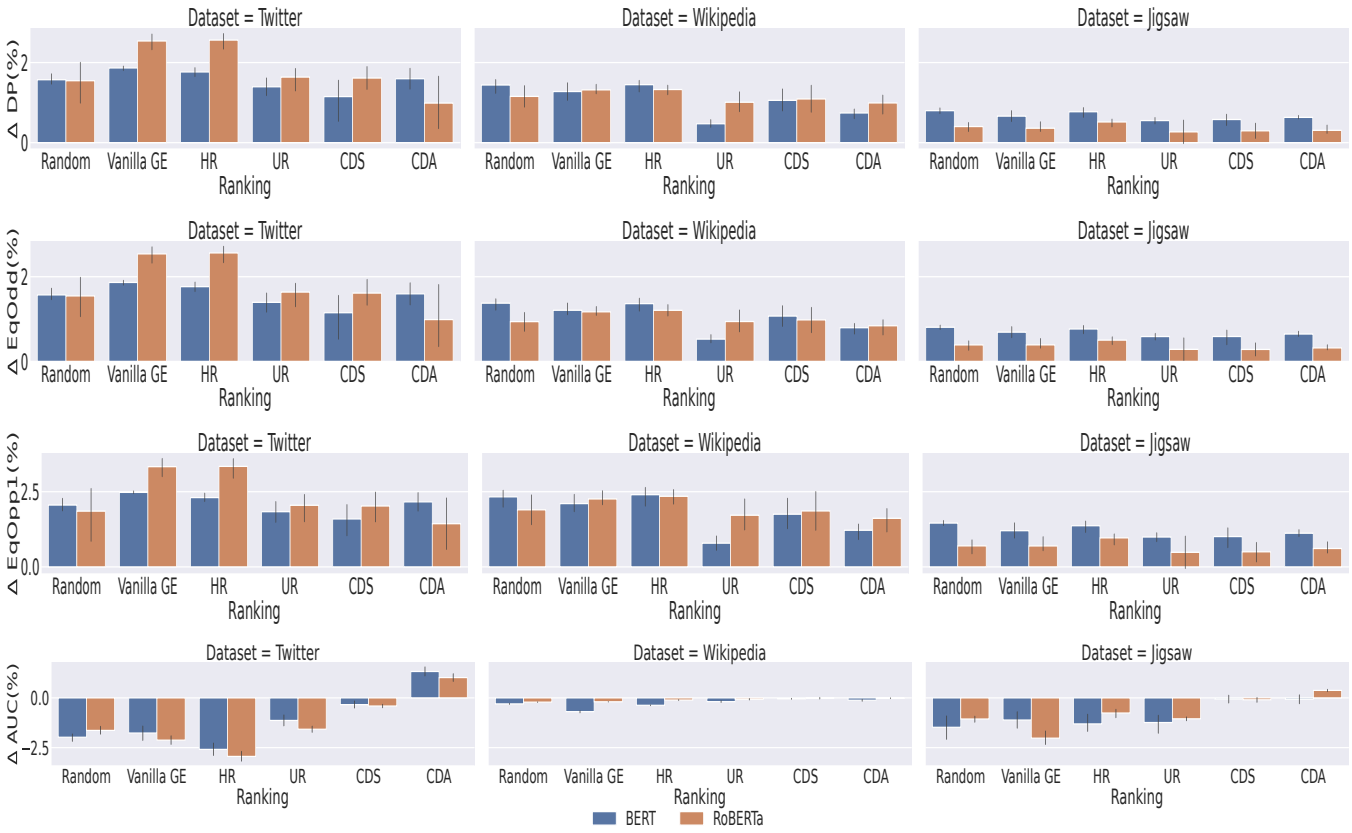

Figure 4: The percentage of change in DP, EqOdd, EqOpp1, and AUC with BERT and RoBERTa on the three datasets using different pruning methods based on the rankings in Table 3, compared to the biased model. HR and UR refer to healthy and unhealthy random rankings, respectively. Both CDA and CDS are added for comparison. For all the fairness metrics, higher values indicate fairer models.

examples based on the descending *GE* score, which we refer to as the "healthy random" ranking. This agrees with the intuition that provided in section 4.2. The fairness obtained by our proposed healthy random ranking outperforms that of CDS and CDA on all the datasets with both BERT and RoBERTa models, using DP, EqOdd, and EqOpp1 metrics. This is accomplished with no more than 3% degradation in the AUC on the downstream task, relative to the model without debiasing. Both the vanilla *GE* and random rankings perform relatively well compared to CDA and CDS, but the healthy random ranking shows more improvement in terms of fairness. It is important to mention that all the ranking methods in Table 3 use no more than half of the examples used by CDA. Note that the AUC for CDA is slightly higher than the baseline on the Twitter dataset because the dataset is related to sexism, so adding more counterfactual examples helps not only in improving fairness but also in improving the performance on the downstream task.

6 Conclusion and Future Work

The intuition behind the reduction in gender bias without a substantial degradation in the performance on different datasets stems from our hypothesis that models can find different ways, *i.e.* different sets of weights, to solve the task. When the model is permitted to learn without any constraints being imposed, it tends to choose the simplest way to solve the given task, which includes learning some undesired associations between a group of features and the output labels (such as thinking that a man is more likely to be a doctor, than a woman). Augmenting the training dataset with counterfactual examples during the fine-tuning step forces the model to *unlearn* these undesired correlations and to find an alternative way to solve the task, which results in a less biased model. However, some of the examples in the augmented dataset can be redundant, or even have an adverse effect through enforcing gender stereotypes that the model had learned during pre-training. Our proposed method that ranks the examples based on the healthy random ranking defined in Table 3 may then be used to prune such examples.

Our future work will be in the direction of generalizing the applicability of our method, such that it includes the bias against other minority groups, such as LGBTQ+, Black people, Jewish people, etc. We believe that extending this work to other languages is also worth exploring.

Acknowledgements

Sarath Chandar is supported by a Canada CIFAR AI Chair and an NSERC Discovery Grant. The authors acknowledge the computational resources provided by Microsoft. We are thankful to Nicolas Le Roux, Marc-Alexandre Côté, Eric Yuan, Andreas Madsen, Romain Laroche, Su Lin Blodgett, and Adam Trischler for their helpful feedback in this project. We are also thankful to the reviewers for their constructive comments.

References

Adi, Y.; Kermany, E.; Belinkov, Y.; Lavi, O.; and Goldberg, Y. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Belinkov, Y.; and Glass, J. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Brunet, M.-E.; Alkalay-Houlihan, C.; Anderson, A.; and Zemel, R. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Conference on Empirical Methods in Natural Language Processing*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS Workshop on Deep Learning*.

Conneau, A.; Kruszewski, G.; Lample, G.; Barrault, L.; and Baroni, M. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Annual Meeting of the Association for Computational Linguistics*.

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

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Dinan, E.; Fan, A.; Wu, L.; Weston, J.; Kiela, D.; and Williams, A. 2020. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Conference on AI, Ethics, and Society*.

Friedman, B.; and Nissenbaum, H. 1996. Bias in computer systems. *Transactions on Information Systems*.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual fairness in text classification through robustness. In *Conference on AI, Ethics, and Society*.

Gupta, U.; Ferber, A.; Dilkina, B.; and Ver Steeg, G. 2021. Controllable Guarantees for Fair Outcomes via Contrastive Information Estimation. In *AAAI Conference on Artificial Intelligence*.

Hall Maudslay, R.; Gonen, H.; Cotterell, R.; and Teufel, S. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*.

Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; and Scholkopf, B. 1998. Support vector machines. *Intelligent Systems and their applications*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hupkes, D.; Veldhoen, S.; and Zuidema, W. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*.

Isabelle, P.; Cherry, C.; and Foster, G. 2017. A Challenge Set Approach to Evaluating Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*.

Jaiswal, A.; Moyer, D.; Ver Steeg, G.; AbdAlmageed, W.; and Natarajan, P. 2020. Invariant representations through adversarial forgetting. In *AAAI Conference on Artificial Intelligence*.

Kiritchenko, S.; and Mohammad, S. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Joint Conference on Lexical and Computational Semantics*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Lu, K.; Mardziel, P.; Wu, F.; Amancharla, P.; and Datta, A. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*.

Manzini, T.; Yao Chong, L.; Black, A. W.; and Tsvetkov, Y. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence

Encoders. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Meade, N.; Poole-Dayan, E.; and Reddy, S. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Annual Meeting of the Association for Computational Linguistics*.

Mordido, G.; and Meinel, C. 2020. Mark-Evaluate: Assessing Language Generation using Population Estimation Methods. In *International Conference on Computational Linguistics*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*.

Naik, A.; Ravichander, A.; Sadeh, N.; Rose, C.; and Neubig, G. 2018. Stress Test Evaluation for Natural Language Inference. In *International Conference on Computational Linguistics*.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. In *Conference on Empirical Methods in Natural Language Processing*.

Paul, M.; Ganguli, S.; and Dziugaite, G. K. 2021. Deep Learning on a Data Diet: Finding Important Examples Early in Training. *Advances in Neural Information Processing Systems*.

Pearl, J. 1995. Causal Diagrams for Empirical Research. *Biometrika*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *OpenAI report*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Conference on Empirical Methods in Natural Language Processing*.

Sauder, J.; Hu, T.; Che, X.; Mordido, G.; Yang, H.; and Meinel, C. 2020. Best Student Forcing: A Simple Training Mechanism in Adversarial Language Generation. In *Language Resources and Evaluation Conference*.

Sennrich, R. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *European Chapter of the Association for Computational Linguistics*.

Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S. 2019. Learning controllable fair representations. In *International Conference on Artificial Intelligence and Statistics*.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Annual Meeting of the Association for Computational Linguistics*.

Toneva, M.; Sordoni, A.; des Combes, R. T.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2019. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *International Conference on Learning Representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *Proc. of the Advances in Neural Information Processing Systems (Neurips)*, 5998–6008.

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL Student Research Workshop*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*.

Zhang, G.; Bai, B.; Zhang, J.; Bai, K.; Zhu, C.; and Zhao, T. 2020. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Annual Meeting of the Association for Computational Linguistics*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics*.