On Bias and Fairness in NLP: Investigating the Impact of Bias and Debiasing in Language Models on the Fairness of Toxicity Detection

Fatma Elsafoury
Fraunhofer Research institute
Weizenbaum Research institute

Stamos Katsigiannis
Durham University

Language models are the new state-of-the-art natural language processing (NLP) models and they are being increasingly used in many NLP tasks. Even though there is evidence that language models are biased, the impact of that bias on the fairness of downstream NLP tasks is still understudied. Furthermore, despite that numerous debiasing methods have been proposed in the literature, the impact of bias removal methods on the fairness of NLP tasks is also understudied. In this work, we investigate three different sources of bias in NLP models, i.e. representation bias, selection bias and overamplification bias, and examine how they impact the fairness of the downstream task of toxicity detection. Moreover, we investigate the impact of removing these biases using different bias removal techniques on the fairness of toxicity detection. Results show strong evidence that downstream sources of bias, especially overamplification bias, are the most impactful types of bias on the fairness of the task of toxicity detection. We also found strong evidence that removing overamplification bias by fine-tuning the language models on a dataset with balanced contextual representations and ratios of positive examples between different identity groups can improve the fairness of the task of toxicity detection. Finally, we build on our findings and introduce a list of guidelines to ensure the fairness of the task of toxicity detection.

1. Introduction

Language models (LM) are being used in different natural language processing (NLP) tasks, like in search engines (Zhu et al. 2023), email filtering (AbdulNabi and Yaseen 2021), and content moderation (Elsafoury et al. 2021). Recent research has indicated that LMs are prone to learning and reproducing harmful social biases (Garg et al. 2018; Caliskan, Bryson, and Narayanan 2017; Sweeney and Najafian 2019; Nangia et al. 2020; Nadeem, Bethke, and Reddy 2021). For example, Elsafoury et al. (2022) demonstrate that different static word embeddings learn to associate between marginalized identities (e.g., Women, LGBTQ, and non-white ethnicity) and profane words. Similar results have been found in contextual word embeddings as well (Ousidhoum et al. 2021; Nozza et al. 2022; Elsafoury 2023). However, the impact of social bias in LMs on downstream NLP tasks like toxicity detection is still understudied. For example, even though it has been shown in the literature that different bias metrics return different results (Badilla, Bravo-Marquez, and Pérez 2020; Elsafoury, Wilson, and Ramzan 2022), most of the studies that investigated the impact of bias in LMs on downstream NLP tasks used only one bias metric (Steed et al. 2022; Goldfarb-Tarrant et al. 2021), which leaves the findings inconclusive.

Understanding the impact of social bias on downstream tasks like toxicity detection is crucial, especially with research demonstrating that content written by marginalized identities is sometimes falsely flagged as toxic or hateful (Sap et al. 2019). Furthermore, various methods have been proposed in the literature for removing bias from LMs and for ensuring that NLP tasks do not

discriminate between different identity groups based on sensitive attributes like gender, ethnicity, religion, or sexual orientation (Zhao et al. 2017; Liang et al. 2020; Webster et al. 2020; Meade, Poole-Dayan, and Reddy 2022; Qian et al. 2022). Yet the impact of these debiasing methods on downstream NLP tasks is also understudied.

In this work, we aim to address these research gaps by investigating the impact of bias and debiasing in LMs on the downstream task of toxicity detection. Lopez (2021) distinguishes between three types of bias: Technological bias, Socio-Technical bias, and Societal bias based on the legal anti-discrimination regulations. In this study, we focus on socio-technical bias which is described as *"A systematic deviation due to structural inequalities."* In particular, we investigate three different sources of socio-technical bias from the predictive bias framework for NLP: (a) *Representation bias*, (b) *Selection bias*, and (c) *Overamplification bias* (Shah, Schwartz, and Hovy 2020; Hovy and Prabhumoye 2021). It must be noted that representation bias is also referred to as *upstream* bias in the literature (Steed et al. 2022), while selection and overamplification bias are refereed to as *downstream* bias (Steed et al. 2022; Badilla, Bravo-Marquez, and Pérez 2020).

We use different metrics from the literature to measure representation bias and fairness to evaluate the consistency of our results. We also propose two metrics to measure selection and overamplification bias. First, we investigate the aforementioned sources of bias and their impact on the fairness of the downstream task of toxicity detection. Then, we remove these sources of bias and investigate whether their removal improves the fairness of toxicity detection. The aim of this work is to find the most impactful sources of bias and the most effective bias removal techniques to use to improve the fairness of toxicity detection without compromising its performance. To this end, this work aims to answer the following research questions:

1. What is the impact of the different sources of bias on the fairness of the downstream task of toxicity detection?
2. What is the impact of removing the different sources of bias on the fairness of the downstream task of toxicity detection?
3. Which debiasing technique to use to ensure the fairness of the task of toxicity detection?
4. How to ensure the fairness of the toxicity detection task?

To answer our research questions, we use statistical definitions of bias and fairness from the NLP literature to measure bias and fairness in three LMs: AlBERT-base-v2 (Lan et al. 2020), BERT-base-uncased (Devlin et al. 2019), and RoBERTa-base (Liu et al. 2019). We measure three sources of bias, i.e. representation bias (§6.1), selection bias (§6.3), overamplification bias (§6.5), using different metrics and investigate their impact on toxicity detection using statistical correlation. Then, we use different methods for representation bias removal (§6.2), selection bias removal (§6.4) and overamplification bias removal (§6.6), and investigate the impact of these bias removal methods on the models' fairness and performance. We then analyse our results to answer the first research question and to understand the impact of the different sources of bias on the models' fairness on toxicity detection (§7.1), as well as to understand the impact of removing different sources of bias on the fairness of toxicity detection and to answer our second research question (§7.2). Thereafter, we further analyze the debiasing (bias removal) results to identify the most effective technique to ensure the models' fairness on the downstream task of toxicity detection and to answer our third research question (§7.3). Finally, to answer the fourth research question, we built on the findings of this work on the fairness of toxicity detection and propose a list of guidelines to ensure the fairness of toxicity detection (§8).

The main contributions of this paper can be summarized as follows:

1. We provide a comprehensive investigation of different sources of bias in NLP models and their impact on the fairness of the task of toxicity detection.

2. We provide a comprehensive investigation of different bias removal (debiasing) methods to remove different sources of bias and their impact on the fairness and performance of the task of toxicity detection.
3. We provide guidelines to ensure the fairness of the task of toxicity detection.

Our findings suggest that the dataset used in measuring fairness impacts the measured fairness scores of toxicity detection, and using a balanced dataset improves the fairness scores. Unlike the findings of previous research (Goldfarb-Tarrant et al. 2021), our findings suggest that upstream (representation) bias is impactful on the fairness of the toxicity detection. However, our results suggest that downstream sources of bias (selection and overamplification) are more impactful on the fairness of the toxicity detection, which is in line with the findings of previous research (Steed et al. 2022). However, unlike the findings of (Steed et al. 2022), our results suggest that removing overamplification bias in the training dataset before fine-tuning is the most effective downstream bias removal method and improved the fairness of the toxicity detection. Our results show strong evidence that the most effective methods to remove overamplification bias is to use the simple method of lexical word replacement to create perturbations to balance the representation of the different identity groups in the training dataset. Finally, we build on these findings and provide a list of guidelines to check against to ensure the fairness of the task of toxicity detection.

Improving the fairness of toxicity detection is very critical for ensuring that the decisions made by the models are not based on sensitive attributes like race or gender. For transparency and to encourage further investigation, we make the code of this work publicly available[1].

2. Background

In this section, we review the literature on the definition of *Bias* and *fairness* and the different metrics used in the literature to measure these two concepts. Then, we provide a critical review of the relevant work in the literature on investigating the impact of bias in LMs on downstream NLP tasks, their limitations, and how our work addresses and mitigates those limitations.

2.1 Bias

The term *bias* is defined and used in many ways, as shown in Olteanu et al. (2019). The normative definition of bias in cognitive science, is: *"Behaving according to some cognitive priors and presumed realities that might not be true at all"* (Garrido-Muñoz et al. 2021). Furthermore, the statistical definition of bias is *"Systematic distortion in the sampled data that compromises its representatives"* (Olteanu et al. 2019). In NLP, while bias has been described in several ways, the statistical definition is the most used in the literature (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018; Nangia et al. 2020; Nadeem, Bethke, and Reddy 2021). In this paper, similar to other works on investigating bias in the literature, we use the statistical definition of bias to measure the socio-technical bias.

Furthermore, in the last few years, various metrics have been proposed in the literature to quantify bias in static word embeddings (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018; Sweeney and Najafian 2019; Dev and Phillips 2019; Elsafoury et al. 2022) and contextual word embeddings (language models) (May et al. 2019; Kurita et al. 2019; Nangia et al. 2020; Nadeem, Bethke, and Reddy 2021; Guo and Caliskan 2021). Other researchers focused on quantifying the NLP models' fairness when used in a downstream task (De-Arteaga et al. 2019; Borkan et al. 2019; Qian et al. 2022), with most of these studies focusing on measuring only representation

[1]This link will be available upon acceptance.

bias in LMs using various proposed metrics, such as CrowS-Pairs (Nangia et al. 2020), StereoSet (Nadeem, Bethke, and Reddy 2021), and SEAT (May et al. 2019).

2.2 Fairness

Fairness has many formal definitions, built on those from literature on the fairness of exam testing from the 1960s, 70s and 80s (Hutchinson and Mitchell 2019). The most recent fairness definitions are broadly categorized into two groups:

- *Individual fairness*, which is defined as "*An algorithm is fair if it gives similar predictions to similar individuals*" (Kusner et al. 2017). Counterfactual fairness is an example of individual fairness metrics and there are different proposed metrics in the literature to measure counterfactual fairness (Kusner et al. 2017; Qian et al. 2022; Krishna et al. 2022; Fryer et al. 2022a).
  For a given model $\hat{Y} : X \rightarrow Y$ with features X, sensitive attributes A, prediction $\hat{Y}$, and two individuals i and j, and if individuals i and j are similar, the model achieves individual fairness if $\hat{Y}(X^i, A^i) \approx \hat{Y}(X^j, A^j)$.
- *Group fairness*, which can be defined as *"An algorithm is fair if the model prediction $\hat{Y}$ and sensitive attribute A are independent"* (Caton and Haas 2020; Kusner et al. 2017). Based on group fairness, the model is fair if $\hat{Y}(X|A = 0) = \hat{Y}(X|A = 1)$. Group fairness is the most common definition used in NLP and there are two main approaches for measuring a model's fairness in that case:
  1. *Threshold-based metrics*, where a model's fairness is measured by how much the classifier's predicted labels differ between different groups of people, based on a threshold (Hutchinson and Mitchell 2019). The equalized odds metric is a threshold-based metric, and it is the most commonly used metric in the literature for measuring fairness in the downstream task of text classification (De-Arteaga et al. 2019; Cao et al. 2022; Steed et al. 2022). Equalized odds are measured by the absolute difference (gap) between the true positive rates (TPR) or false positive rates (FPR) between different groups of people, g and $\hat{g}$, based on sensitive attributes like gender, race, etc, as defined in Equations 1Fairnessequation.0.2.1 and 2Fairnessequation.0.2.2.

$$FPR\_gap_{g,\hat{g}} = |FPR_g - FPR_{\hat{g}}| \tag{1}$$

$$TPR\_gap_{g,\hat{g}} = |TPR_g - TPR_{\hat{g}}| \tag{2}$$

  2. *Threshold-agnostic metrics*, where fairness is measured by how much the distribution of the classifier's prediction probabilities varies across different groups of people based on their sensitive attributes. Borkan et al. (2019) propose a set of fairness metrics that are based on the AUC score. An important advantage of the threshold-based metrics is that they are robust to data imbalances in the amount of positive and negative examples in the test set (Borkan et al. 2019). In this work, we measure the absolute difference in the area under the curve (AUC) scores between marginalized group (g) and non-marginalized group $\hat{g}$, as shown in Equation 3Fairnessequation.0.2.3.

$$AUC\_gap_{g,\hat{g}} = |AUC_g - AUC_{\hat{g}}| \tag{3}$$

Nevertheless, the current metrics used to measure bias and fairness in NLP have some limitations. For example, there is a lack of articulation of what the different metrics actually measure, and ambiguities and unstated assumptions, as discussed in Blodgett et al. (2021). There are also limitations with using sentence templates to measure the bias and fairness of LMs on downstream tasks. For example, Seshadri, Pezeshkpour, and Singh (2022) demonstrate that fairness scores vary with small lexical changes in the sentence templates that do not change the semantics of the sentences. However, these metrics could still be used as initial indicators of the Socio-technical bias in LMs and downstream NLP tasks. In this work, we use these metrics in that capacity as initial indicators of bias in LMs and of how they impact the fairness of toxicity detection. To mitigate for these limitations, we start our investigation using different group fairness metrics. Then, we use individual fairness metrics to further confirm the results of the group fairness metrics.

Furthermore, the focus of studying bias in the literature has been on LMs trained on English data and from a Western perspective where the marginalized identities tend to be women, non-white ethnicities, and minority religious groups like Muslims and Jews. Similarly, due to the datasets used in our paper, we focus our analysis on LMs trained on English data, and we investigate bias from a Western-perspective and focus our analysis on three sensitive attributes: Gender, Race, and Religion.

2.3 Related work

The impact of bias on models' fairness on NLP downstream tasks is understudied. The majority of studies have focused on the impact of representation bias (Cao et al. 2022; Kaneko, Bollegala, and Okazaki 2022; Goldfarb-Tarrant et al. 2021), excluding the impact of other sources of bias like selection and overamplification. Some researchers found no strong evidence that representation bias in LM impacts fairness in downstream NLP tasks (Cao et al. 2022; Kaneko, Bollegala, and Okazaki 2022). Steed et al. (2022) found a positive correlation between representation bias and fairness of downstream tasks but argued that this positive correlation is a result of cultural artifacts found in both pre-training and fine-tuning datasets.

However, there are some limitations to those studies. For example, in Steed et al. (2022), the authors use two representation bias metrics which use bleached template sentences, which are sentences that don't have a real semantic context, to measure bias, these metrics have been criticized as they may not be semantically bleached (May et al. 2019). Since the used bias metrics are not reliable, their impact on the fairness of NLP tasks is also unreliable. Moreover, both Cao et al. (2022) and Steed et al. (2022) use different representation bias metrics for the two text classification tasks examined, which results in a lack of consistency in their findings.

Similarly, the impact of removing bias on the models' fairness in downstream tasks has been understudied. For example, Meade, Poole-Dayan, and Reddy (2022) investigate the impact that different debiasing approaches have on the performance of different NLP downstream tasks. However, they don't investigate the impact debiasing has on the fairness of the downstream tasks. Baldini et al. (2022) investigate the impact of removing bias from fine-tuned models, excluding the impact of removing representation bias on the fairness of downstream tasks.

On the other hand, in Kaneko, Bollegala, and Okazaki (2022), the authors investigate the effectiveness of different debiasing methods that remove different sources of bias representation bias and overamplification bias on the fairness of the downstream tasks. However, the authors do not investigate the impact of removing selection bias on the fairness of downstream NLP tasks. Moreover, for removing overamplification bias, the authors' investigation covers only gender bias in the downstream tasks of occupation classification, semantic textual similarities (STS) and natural language inference (NLI), excluding other bias types like racial and religion bias and tasks like toxicity detection.

Similarly, Steed et al. (2022) investigate the impact of different bias removal techniques to remove representation and selection bias on the fairness of downstream NLP tasks. However, the authors don't investigate the impact of removing overamplification bias on the fairness of downstream NLP tasks. Additionally, the authors apply the selection bias removal investigation only on gender bias and only for the task of occupation classification.

As for measuring models' fairness on downstream tasks, Goldfarb-Tarrant et al. (2021) measure fairness as performance gap between different identity groups. However, more recently, fairness is measured as the gap in the false or true positive rates between the different identity groups (De-Arteaga et al. 2019; Steed et al. 2022; Baldini et al. 2022; Kaneko, Bollegala, and Okazaki 2022).

On the other hand, Cao et al. (2022); Kaneko, Bollegala, and Okazaki (2022); Steed et al. (2022); Baldini et al. (2022), use only threshold-based fairness bias metrics for the text classification task. For example, Cao et al. (2022); Kaneko, Bollegala, and Okazaki (2022) use FPR gap to measure fairness on toxicity detection. Similarly, Steed et al. (2022) use TPR gap to measure fairness in the task of occupation classification and FPR gap for the task of toxicity detection. However, more recent research demonstrate that different group fairness metrics give different results (Jourdan et al. 2023) .

Threshold-agnostic metrics have not been widely used to measure fairness and to investigate its correlation to representation bias. Even though, according to Borkan et al. (2019), threshold-agnostic metrics can capture the behavior of the model. Similarly, individual fairness metrics, e.g., counterfactual fairness, have not been used in investigating the impact of bias. Even though, counterfactual fairness provides an in-depth analysis of how the model treats different sentences based on the identity group present in the sentences. Moreover, in most of the studies that investigate the fairness of the task of toxicity detection, the authors don't explain how they measure the fairness of the models between the different identity groups (Cao et al. 2022; Steed et al. 2022). Other researchers do not measure fairness as a notion of discrimination between marginalized and non- marginalized groups but rather as the mention of sensitive topics. For example, Baldini et al. (2022) measure fairness using "*coarse-grained groups (e.g., mention of any religion) instead of the fine-grained annotations (e.g., Muslim)*". Since the authors do not actually measure fairness but rather how a model treats the existence of a sensitive topic, the findings of their work might not be reliable or extendable to fairness as a notion of discrimination between different identity groups.

This paper fills these research gaps in the literature by investigating different sources of bias and their impact on the models' fairness in the downstream task of toxicity detection. We investigate bias and fairness for different sensitive attributes: gender, race, and religion. We aim to overcome the limitations of previous research by using different metrics to measure representation bias and models' fairness. Moreover, we investigate the effectiveness of various bias mitigation methods (debiasing) for removing different sources of bias: representation, selection and overamplification, as well as their impact on the fairness and performance of the task of toxicity detection.

3. Methodology

Four groups of experiments were conducted to investigate the impact of each source of bias on the fairness of the downstream task of toxicity detection. Figure 1Overview of conducted investigation.figure.caption.1 provides an overview of these four groups of experiments. As shown in Figure 1Overview of conducted investigation.figure.caption.1, in Step 1, we first measured the fairness of the toxicity detection task (§5) using various models and used these fairness scores as a baseline. Then, in Step 2A, we measured the representation bias in the inspected models and its impact on the models' fairness on the task of toxicity detection (§6.1), as well as the

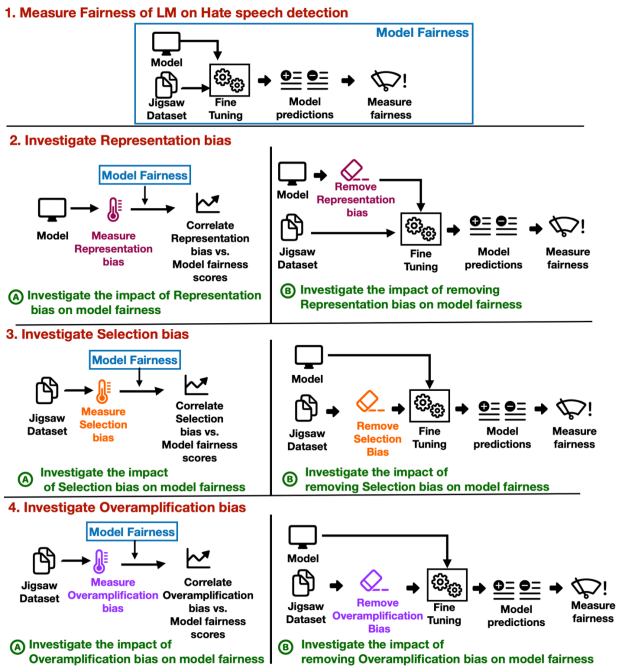

Figure 1: Overview of conducted investigation.

impact of removing representation bias (§6.2), as shown in Step 2B of Figure 1Overview of conducted investigation.figure.caption.1. In Steps 3 (A & B) and 4 (A & B), we repeated the same investigation for selection bias (§6.3 & §6.4) and overamplification bias (§6.5 & §6.6). Then, we investigated the impact of removing multiple sources of biases on the fairness of the task of toxicity detection (§6.7). Finally, we build on these findings and recommend guidelines to achieve fairer toxicity detection (§8).

4. Toxicity detection

4.1 Dataset

We use the civil community dataset (Borkan et al. 2019) in our experiments. The dataset contains almost 2 million comments from the Civil Comments Platform, labelled as toxic or not, along with labels on the identity of the target of the sentence, e.g., religion, sexual orientation, gender, and race. The identity labels provided in the dataset are both crowd-sourced and automatically labelled. When we analyzed the dataset, we found some issues with the identity labels, e.g., some data items are labelled to contain more than one identity (male and female) as the target of the toxicity[1]. Then, the dataset was pre-processed to keep only the data items where the identity information is labelled by human annotators. Additionally, we follow the same data pre-processing steps used in Elsafoury et al. (2021), where the authors train a BERT model for the task of cyberbullying detection. To this end, we remove URLs and non-ASCII characters, lowercase all the letters, convert all contractions to their formal format and add a space between words and punctuation

[1] https://huggingface.co/datasets/google/civil_comments

Sensitive attribute	Marginalized	Non-marginalized
Gender	Female	Male
Race	Black, Asian	White
Religion	Jewish, Muslim	Christian

Table 1: The examined sensitive attributes and identity groups.

marks. This resulted in 400K data items. The dataset is then split into 40% training, 30% validation, and 30% test sets. We limit this investigation to 3 sensitive attributes, i.e., gender, religion, and race, as shown in Table 1The examined sensitive attributes and identity groups.table.0.1.

We only use the civil community dataset because, to the best of our knowledge, it is the only available toxicity dataset that contains information on both marginalized and non-marginalized identities, which is important to the way we measure fairness, as explained in section 5Fairness in the task of toxicity detectionsection.0.5. Other datasets, like ToxiGen (Hartvigsen et al. 2022), SocialFrame (Sap et al. 2020), Ethos (Mollas et al. 2022), and the MLM data (Ousidhoum et al. 2019) contain information only about marginalized groups, and thus cannot be used in our investigation. The HateXplain dataset (Mathew et al. 2021) contains information about both marginalized and non-marginalized identities. However, HateXplain uses offensive words to refer to marginalized groups, e.g., n*gger to refer to Black people, and identity words to describe non-marginalized groups, e.g., white to refer to Caucasians. This makes the HateXplain dataset unsuitable for the experiments held in section 5Fairness in the task of toxicity detectionsection.0.5 where we create data perturbations, since replacing an offensive word that describes marginalized groups with an identity word to describe a non-marginalized identity group changes the meaning of the sentence and hence its label as hateful or not. Conversely, in the civil community dataset, identity words are used to describe both marginalized and non-marginalized groups. Even the toxic sentences in the civil community dataset do not contain offensive words to describe marginalized groups. This is evident in the most common nouns and adjectives found in the toxic and non-toxic sentences that are targeted at the different identity groups, as shown in Table 2The most common ten adjectives and nouns as generated by the Spacy Python package in the toxic and non-toxic sentences that are targeted at different identity groups in the civil community fairness dataset.table.0.2. It must be noted that the most common nouns and adjectives were generated using the Spacy Python package[1].

4.2 Language models

The fairness of the downstream task of toxicity detection is evaluated on the following widely used models: BERT-base-uncased (Devlin et al. 2019), RoBERTa-base (Liu et al. 2019), and ALBERT-base (Lan et al. 2020). We fine-tune these models on the civil community dataset. Following the experimental setting from (Elsafoury et al. 2021), the models are fine-tuned for 3 epochs, using a batch size of 32, a learning rate of $2e^{-5}$, and a maximum text length of 61. Classification results using the fine-tuned models indicate that ALBERT-base is the best-performing model, with an AUC score of 0.911, followed by RoBERTa-base with an AUC score of 0.908, and BERT-base with an AUC score of 0.902. The fine-tuned models are then used to measure fairness in the toxicity detection task.

[1] `https://spacy.io/api`

Identity	Toxic sentences	Non-toxic sentences
Black	black, people, blacks, racist, police,Black, other, man, white, men	black, people, blacks, man, police,other, Black, white, many, men
Asian	people, Asian, many, repair, chef, country, racist, real, citizens, Korean	Asian, other, Chinese, people, many, countries, years, women, more, country
White	white, people, racist, men, supremacists, man, racism, right, supremacist, White	white, people, men, racist, right, other, man, many, supremacists, male
Female	women, woman, people, white, other, many, sexual, time, life, sex	women, woman, people, many, other, more, time, right, life, abortion
Male	man, men, white, black, people, male, women, stupid, racist, males	man, men, white, people, male, other, many, right, time, way
Muslim	Muslim, people, women, white, many, other, muslim, terrorists, religion, muslims	Muslim, people, countries, women, other, many, country, ban, world, muslim
Jewish	Jewish, people, anti, black, hate, women, good, other, white, man	Jewish, people, anti, other, white, right, way, state, many, world
Christian	people, white, Christian, women, right, other, many, sex, Catholic,life	Catholic, people, Christian, church, many, women, other, right, time, good

Table 2: The most common ten adjectives and nouns as generated by the Spacy Python package in the toxic and non-toxic sentences that are targeted at different identity groups in the civil community fairness dataset.

5. Fairness in the task of toxicity detection

To evaluate the fairness of the examined models on the downstream task of toxicity detection, we used two sets of fairness metrics: (i) Threshold-based, which uses the absolute difference (gap) in the false positive rates (FPR) and true positive rates (TPR) between the marginalized group (g) and non-marginalized group $\hat{g}$, as shown in Equations 1Fairnessequation.0.2.1 and 2Fairnessequation.0.2.2, and (ii) Threshold-agnostic metrics, which measure the absolute difference in the area under the curve (AUC) scores between marginalized group (g) and non-marginalized group $\hat{g}$, as shown in Equation 3Fairnessequation.0.2.3. In cases where there is more than one identity group in the marginalized group for a sensitive attribute, e.g., Asian and Black vs. White, we measured the mean of the FPR, TPR, and AUC scores of the two groups, e.g., Asian and Black, and then use that score to represent the marginalized group (g). These scores express the amount of unfairness in the toxicity detection models, with higher scores denoting less fair models and lower scores denoting fairer models. These fairness metrics are measured between two groups, marginalized and non-marginalized, similar to the approach used by Elsafoury et al. (2022).

5.1 Balanced fairness dataset

To measure fairness, we filtered the test set to ensure that the sentences contain only one identity group, which resulted in 21K sentences to improve the quality of the measured fairness. We found differences in the sizes of the subsets of sentences that mention the different identity groups. For example, the size of the subset of sentences that are targeted at males is 3,716 sentences while the size of the female subset is 6,046 sentences. We also found differences in the ratio of the positive samples between the different identity groups that belong to the same sensitive attribute. The ratio of positive samples for the male and female groups are 0.12 and 0.10 respectively; for the White, Asian, and Black groups are 0.20, 0.07, and 0.27 respectively; and for the Christian, Muslim, and Jewish groups are 0.05, 0.16, and 0.12 respectively. We hypothesize that these differences between the different identity groups may influence the fairness scores.

To test this hypothesis, we created a balanced fairness dataset and used it to measure fairness in the fine-tuned models. To create this balanced fairness dataset, we followed previous research and augmented our dataset using perturbations (Webster et al. 2020). Previous studies, like Garg et al. (2019) and (Fryer et al. 2022b), used data perturbations to improve the fairness of the task of toxicity detection. However, the authors of these two studies chose to create the data perturbations only on the non-toxic sentences to avoid what Garg et al. (2019) describe as *Asymmetric Counterfactuals*. Garg et al. (2019) argue that asymmetric counterfactuals happen in toxic sentences for two reasons:

1. When the toxicity targets a marginalized group, it is based only on the identity, and no other toxicity signals are present. In such cases substituting identities would result in non-toxic sentence and hence equal prediction should not be required over most counterfactuals.
2. Stereotyping comments are more likely to occur in a toxic comment attacking the marginalized group than in a non-toxic comment referencing some other identity group.

However, the results in Table 2The most common ten adjectives and nouns as generated by the Spacy Python package in the toxic and non-toxic sentences that are targeted at different identity groups in the civil community fairness dataset.table.0.2 show that the most common nouns and adjectives in the toxic and non-toxic sentences that target marginalized and non-marginalized groups are almost similar. For example, words like “black”, “blacks”, “police” are used in both toxic and non-toxic comments to describe Black people and words like “white”, “racist”, “supremacists” exist in both toxic and non-toxic sentences that describe White people. This challenges the hypothesis made by Garg et al. (2019) that stereotypes are more likely to happen in toxic sentences. Additionally, the offensive words found in the toxic comments targeted at marginalized groups are not based on identity. For example, Table 2The most common ten adjectives and nouns as generated by the Spacy Python package in the toxic and non-toxic sentences that are targeted at different identity groups in the civil community fairness dataset.table.0.2 shows that the word “racist” is used in the toxic comments targeted at White, Black, and Asian people. This challenges the second hypothesis made by Garg et al. (2019) that toxicity targeted at marginalized identities is based mainly on the identity with no other toxicity signal. These findings suggest that in our dataset, asymmetric counterfactuals are not more likely to happen in toxic sentences than non-toxic sentences. Hence and based on this evidence, in this work, we decided to create perturbations for both toxic and non-toxic sentences.

To this end, lexical word replacement was used to create perturbations of existing sentences using regular expressions. As explained earlier, this is possible with the civil community dataset because after inspecting the most common nouns and adjectives used in each subset that targets a certain identity, it was found that the most common words are non-offensive words that describe that identity. For example, among the most common nouns in the data subset that are targeted at black people are: “black” and “blacks”. The most common nouns in the data subset targeted

at Asian people are “asian” and “chinese”. A similar pattern is found for religion and gender identities. However, this approach is not suitable for gender perturbations, as, in the English language, pronouns also change between males and females. To this end, perturbations for the male and female identity groups are created using the AugLy[1] tool, which is provided by Facebook research to swap gender information (Papakipos and Bitton 2022).

The balanced fairness dataset contains 55,476 samples and has the same ratio between positive and negative samples for each identity group within the same sensitive attribute. For example, the ratio of the positive (toxic) examples in the male and female identity groups is 0.10 for the gender attribute, the ratio of the positive samples for the Black, White and Asian groups is 0.20 for the race attribute, and the ratio of the positive samples for the Muslim, Christian and Jewish groups is 0.10 for the religion attribute.

5.2 Fairness results

The fairness evaluation results for the three examined models on the original and the balanced civil community fairness datasets are shown in Table 3The fairness scores of the examined models on the original and the balanced civil community fairness datasets. (↑) denotes that the fairness score increased, and the fairness worsened. (↓) denotes that the fairness score decreased, and the fairness improved.table.0.3. From this table, it is evident that the use of the balanced civil community fairness dataset led to improved fairness scores across most of the fairness metrics and across all models. This finding suggests that the dataset used to measure the fairness in the downstream task of toxicity detection impacts the measured fairness, and it is important to ensure that there is a balanced representation of the different identity groups to get reliable fairness scores. This finding supports our hypothesis and, to the best of our knowledge, has not been mentioned in the literature on measuring fairness before.

The results also indicate that when we use the original imbalanced fairness dataset, the different metrics used to measure the fairness reported different fairness scores related to each sensitive attribute in the fine-tuned models. We used Pearson’s correlation coefficient (ρ) to measure how different the fairness metrics are and found that even though the FPR_gap and TPR_gap are both threshold-based metrics, there is no positive correlation between the two metrics for the three models. There is a negative correlation between TPR_gap and FPR_gap (ρ = -0.37), a negative correlation between the TPR_gap and the AUC_gap scores for the three models (ρ = -0.42), and a positive correlation between the FPR_gap and the AUC_gap for the models (ρ = 0.46).

On the other hand, when we used the balanced civil community fairness dataset to measure fairness, we found a positive correlation between all the fairness metrics in all three models. There is a positive correlation between the FPR_gap and the TPR_gap scores (ρ = 0.59), a positive correlation between FPR_gap and AUC_gap scores (ρ = 0.64), and a positive correlation between the TPR_gap and the AUC_gap scores (ρ = 0.27). This is another evidence that using a fairness dataset with a balanced representation of the different identity groups leads to more reliable fairness scores. For the rest of this work, the balanced fairness dataset is used to measure fairness. Similarly, in the rest of the paper, to investigate the impact of the different sources of bias in the next sections, we use Pearson’s correlation between the bias scores and the fairness scores measured in this section similar to the work done in Steed et al. (2022); Kaneko, Bollegala, and Okazaki (2022); Cao et al. (2022).

[1] https://augly.readthedocs.io/en/latest/README.html

Attribute	Model	Dataset	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	Original	0.001	0.081	0.025
		Balanced	↑ 0.006	↓ 0.038	↓ 0.003
	BERT	Original	0.002	0.111	0.026
		Balanced	↑ 0.008	↓ 0.036	↓ 0.009
	RoBERTa	Original	0.007	0.084	0.017
		Balanced	↓ 0.004	↓ 0.031	↓ 0.011
Race	ALBERT	Original	0.007	0.044	0.003
		Balanced	↑ 0.008	↓ 0.0016	↑ 0.018
	BERT	Original	0.008	0.017	0.048
		Balanced	↑ 0.015	↓ 0.002	↓ 0.025
	RoBERTa	Original	0.014	0.127	0.028
		Balanced	↓ 0.003	↓ 0.011	↓ 0.021
Religion	ALBERT	Original	0.019	0.060	0.042
		Balanced	↓ 0.009	↑ 0.108	↓ 0.020
	BERT	Original	0.016	0.027	0.051
		Balanced	↓ 0.008	↑ 0.062	↓ 0.012
	RoBERTa	Original	0.027	0.030	0.0369
		Balanced	↓ 0.021	↑ 0.160	↓ 0.027

Table 3: The fairness scores of the examined models on the original and the balanced civil community fairness datasets. (↑) denotes that the fairness score increased, and the fairness worsened. (↓) denotes that the fairness score decreased, and the fairness improved.

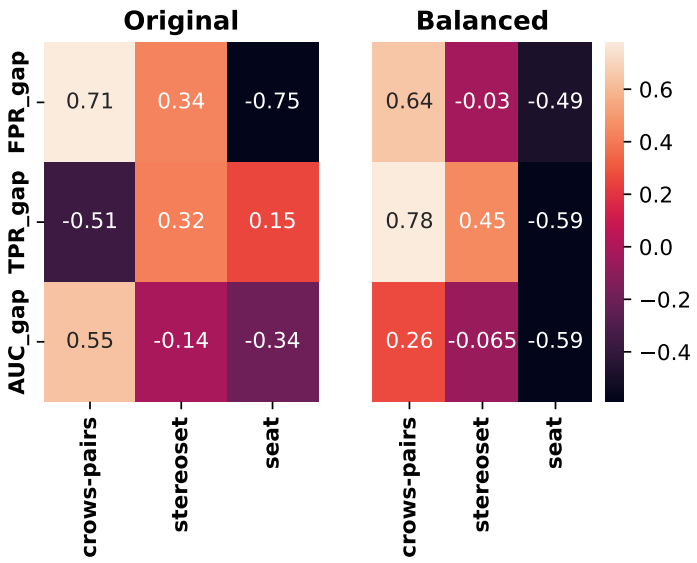

Figure 2: Heatmap of Pearson's correlation between representation bias scores of all LMs and fairness scores of LMs on the downstream task of toxicity detection, on the original civil community fairness dataset (left) and the balanced civil community fairness dataset (right), for all the sensitive attributes.

6. Sources of bias

Shah, Schwartz, and Hovy (2020) present a framework for bias in NLP which includes four sources of bias in NLP models that impact a model's outcome, which are representation bias, label bias, selection bias, and overamplification bias. In this work, we examine how the following three sources of bias impact the examined models' fairness on the task of toxicity detection: (i) representation bias, (ii) selection bias (also called sample bias), and (iii) overamplification bias. We do not investigate label bias because there is no information available on the annotators of the examined civil community dataset. Furthermore, we removed each source of bias and investigated the impact of the bias removal on the fairness of toxicity detection.

6.1 Representation bias

Representation bias describes the societal stereotypes that language models encoded during pre-training. We used three metrics to measure representation bias: CrowS-Pairs (Nangia et al. 2020), StereoSet (Nadeem, Bethke, and Reddy 2021), and SEAT (May et al. 2019) to measure three types of social bias: gender, religion, and race. As shown in Table 4Representation bias scores in the examined models using different bias metrics before and after removing bias using the SentDebias algorithm. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased, and the fairness improved.table.0.4, the results indicate that RoBERTa is the most biased model according to CrowS-Pairs and StereoSet.

We then investigated the impact of representation bias in the inspected models, BERT, ALBERT, and RoBERTa, on their fairness on the task of toxicity detection. To measure that impact, we measure the Pearson's correlation coefficient (ρ) between fairness scores measured by the different fairness metrics and the representation bias scores measured by the different representation bias metrics (Figure 2Heatmap of Pearson's correlation between representation bias scores of all LMs and fairness scores of LMs on the downstream task of toxicity detection, on the original civil community fairness dataset (left) and the balanced civil community fairness dataset (right), for all the sensitive attributes.figure.caption.2 (right)). We found a consistent positive correlation between the CrowS-Pairs bias scores with the fairness scores measured by all three fairness metrics (FPR_gap. TPR_gap and AUC_gap) for all the models and sensitive attributes. On the other hand, there is an inconsistent correlation with the StereoSet scores. This finding is different from previous research that suggested that there is no correlation between representation bias and fairness scores (Cao et al. 2022; Kaneko, Bollegala, and Okazaki 2022). We hypothesize that previous research did not use a balanced fairness dataset, which is why they did not find a consistent positive correlation with fairness metrics. To test this hypothesis, we measured the Pearson correlation coefficient (ρ) between representation bias scores and fairness scores measured using the different fairness metrics on the original imbalanced fairness dataset. We found no consistent positive correlation between representation bias and fairness scores, which supports our hypothesis as shown in Figure 2Heatmap of Pearson's correlation between representation bias scores of all LMs and fairness scores of LMs on the downstream task of toxicity detection, on the original civil community fairness dataset (left) and the balanced civil community fairness dataset (right), for all the sensitive attributes.figure.caption.2 (left). Another reason why previous research has not found a consistent positive correlation is that they used only one metric for either representation bias or fairness. However, using more than one metric helped to reveal this correlation.

Model	CrowsPairs			StereoSet			SEAT		
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
AlBERT-base	0.541	0.513	0.590	0.599	0.575	0.603	0.622	0.551	0.430
+ SentDebias-gender	↓ 0.461	↓ 0.436	↓ 0.466	↓ 0.517	↓ 0.552	↓ 0.586	0.622	0.551	0.430
+ SentDebias-race	↑ 0.564	↓ 0.440	↑ 0.666	↓ 0.542	↓ 0.521	↓ 0.555	0.622	0.551	0.430
+ SentDebias-religion	↑ 0.549	↑ 0.660	↓ 0.581	↓ 0.501	↓ 0.529	↓ 0.510	0.622	0.551	0.430
BERT-base-uncased	0.580	0.581	0.714	0.607	0.5702	0.597	0.620	0.620	0.491
+ SentDebias-gender	↓ 0.427	↓ 0.555	↓ 0.647	↓ 0.475	↓ 0.476	↓ 0.504	0.620	0.620	0.491
+ SentDebias-race	↓ 0.534	↓ 0.398	↓ 0.704	↓ 0.467	↓ 0.562	↓ 0.489	0.620	0.620	0.491
+ SentDebias-religion	↓ 0.534	↑ 0.675	↓ 0.561	↓ 0.469	↓ 0.511	↓ 0.399	0.620	0.620	0.491
RoBERTa-base	0.606	0.527	0.771	0.663	0.616	0.642	0.939	0.307	0.126
+ SentDebias-gender	↓ 0.467	↑ 0.691	↓ 0.561	↓ 0.518	↓ 0.497	↓ 0.477	0.939	0.307	0.126
+ SentDebias-race	↓ 0.429	↓ 0.467	↓ 0.419	↓ 0.485	↓ 0.488	↓ 0.486	0.939	0.307	0.126
+ SentDebias-religion	↓ 0.413	↓ 0.478	↓ 0.352	↓ 0.516	↓ 0.497	↓ 0.486	0.939	0.307	0.126

Table 4: Representation bias scores in the examined models using different bias metrics before and after removing bias using the SentDebias algorithm. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased, and the fairness improved.

6.2 Representation bias removal

We used the SentDebias (Liang et al. 2020) to remove different types of bias from the LMs by projecting a sentence representation onto the estimated bias subspace and subtracting the resulting projection from the original sentence representation. The authors in Liang et al. (2020) compute the bias subspace by following these steps:

1. They define a list of identity words e.g., “woman/man”.
2. They contextualize the identity words into sentences by finding sentences that contain those identity words in public datasets like SST [1] and WikiText-2 [2].
3. They obtain the representation of the contextualized sentence from the pre-trained LM.
4. Principle Component Analysis (PCA) (Abdi and Williams 2010) then used to estimate principal directions of variations of the sentences’ representations. The first K principal components are taken to define the bias subspace.

We used SentDebias to remove gender, racial, and religious bias from the inspected models using the same code and experimental settings shared by Meade, Poole-Dayan, and Reddy (2022). SentDebias seems to reduce the bias scores, as shown in Table 4Representation bias scores in the examined models using different bias metrics before and after removing bias using the SentDebias algorithm. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased, and the fairness improved.table.0.4, in all the models according to the StereoSet and CrowS-Pairs metrics. On the other hand, the SEAT metric, did not show any difference in the bias scores for the debiased models, unlike the reported results in Meade, Poole-Dayan, and Reddy (2022). We found that, according to some metrics, removing one type of bias sometimes leads to exacerbating another type of bias. For example, in Table 4Representation bias scores in the examined models using different bias metrics before

[1] https://huggingface.co/datasets/sst
[2] https://huggingface.co/datasets/mindchain/wikitext2

and after removing bias using the SentDebias algorithm. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased, and the fairness improved.table.0.4, removing racial bias increased the gender bias and removing religion bias increased racial and gender bias in AlBERT-base according to the CrowS-Pairs metric. The same finding is reported in Elsafoury et al. (2022) for static word embeddings.

Furthermore, we investigate the impact of removing representation bias on the models' fairness. We fine-tuned BERT-base, ALBERT-base and RoBERTa-base after debiasing them to remove gender, racial and religious bias. Then, we measure the fairness of the models using the different fairness metrics, threshold-based and threshold-agnostic metrics. The results in Table 5Toxicity detection performance and fairness scores for all models before and after removing representation bias using SentDebias. **Bold** values refer to better AUC scores and better performance on toxicity detection. (↑) denotes that the fairness metric score increased, and the fairness worsened. (↓) denotes that the fairness metric score decreased and the fairness improved. The word *upstream* is used here to refer to removing bias from the models before fine-tuning them on the task of toxicity detection.table.0.5 indicate that, in some cases, removing representation bias, worsened the performance of the models slightly. In other cases, removing representation bias also improved the performance slightly. For example, removing gender bias information increased slightly the AUC scores, in BERT and RoBERTa. This is because the debiased models tend to predict more positive class examples, leading to more true positives and more false positives.

To simplify the analysis of the results, we investigated the impact of removing a certain type of bias on the fairness of the matching sensitive attribute. For example, we analyzed the impact of removing gender bias from the model representation (+ Upstream-SentDebias-gender) on the fairness of the models regarding the gender-sensitive attribute. For most of the models, the majority of the fairness metrics show that removing a certain type of bias from the model representation using SentDebias did not improve fairness for the corresponding sensitive attribute. There are improvements according to certain metrics, but these improvements are inconsistent across sensitive attributes and models. As for the cases where all fairness metrics show improvement in fairness, we find that only removing religion bias from RoBERTa-base representations improved the models' fairness for the religion sensitive attributes in the downstream task of toxicity detection. On the other hand, sometimes removing different types of bias from the model representation led to improvement in fairness for a different sensitive attribute. For example, in the BERT model, according to the FPR_gap and TPR_gap metric, removing racial bias information from the model's representation improved the fairness for the gender-sensitive attributes. Yet again, these findings are inconsistent across all models and across all fairness metrics.

These results suggest that even though there is a positive correlation between representation bias in the inspected models and the models' fairness on the task of toxicity detection bias scores, removing representation bias did not consistently improve the models' fairness. Similar findings were made by (Kaneko, Bollegala, and Okazaki 2022). This could be because the current measures used to remove representation bias are superficial, as argued by Gonen and Goldberg (2019).

6.3 Selection bias

Selection bias (Hovy and Prabhumoye 2021), also known as sample bias, is a result of non-representative observations in the training datasets used in downstream tasks (Shah, Schwartz, and Hovy 2020). For the task of toxicity detection, we interpreted selection bias as the over-representation of a certain identity group with the positive (toxic) label, as shown in Figure 3The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after re-stratification (right).figure.caption.3 (Left). We measure selection bias in the civil community training dataset by measuring the difference in the

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	**0.847**	0.006	0.039	0.004
	+ upstream-sentDebias-gender	0.840	0.006	↓ 0.032	0.004
	BERT	0.830	0.090	0.036	0.010
	+ upstream-sentDebias-gender	**0.841**	↓ 0.011	↑ 0.049	↓ 0.006
	RoBERTa	0.851	0.005	0.032	0.011
	+ upstream-sentDebias-gender	**0.856**	↑ 0.006	↓ 0.022	↓ 0.003
Race	ALBERT	**0.847**	0.008	0.002	0.019
	+ upstream-sentDebias-race	0.838	↓ 0.003	↑ 0.003	↓ 0.013
	BERT	**0.830**	0.016	0.002	0.026
	+ upstream-sentDebias-race	0.829	↑ 0.021	↑ 0.005	↓ 0.024
	RoBERTa	0.851	0.003	0.011	0.021
	+ upstream-sentDebias-race	**0.854**	↑ 0.017	↓ 0.009	0.021
Religion	ALBERT	**0.847**	0.010	0.109	0.020
	+ upstream-sentDebias-religion	0.837	↑ 0.019	↓ 0.094	↓ 0.016
	BERT	0.830	0.008	0.063	0.012
	+ upstream-sentDebias-religion	**0.833**	↑ .015	↑ 0.084	↑ 0.017
	RoBERTa	**0.851**	0.022	0.160	0.027
	+ upstream-sentDebias-religion	0.843	↓ 0.021	↓ 0.100	↓ 0.003

Table 5: Toxicity detection performance and fairness scores for all models before and after removing representation bias using SentDebias. **Bold** values refer to better AUC scores and better performance on toxicity detection. (↑) denotes that the fairness metric score increased, and the fairness worsened. (↓) denotes that the fairness metric score decreased and the fairness improved. The word *upstream* is used here to refer to removing bias from the models before fine-tuning them on the task of toxicity detection.

ratios of the positive examples, between the marginalized and non-marginalized groups. Equation 4Selection biasequation.0.6.4 shows how we to measure selection bias ($Selection_{g,\hat{g}}$) where (N_g) is the size of the data subset that is targeted at marginalized groups (g); ($N_{\hat{g}}$) is the size of the data subset that is targeted at non-marginalized groups ($\hat{g}$); ($N_{g,toxicity=1}$) is the number of toxic sentences that are targeted at marginalized groups; and ($N_{\hat{g},toxicity=1}$) is the number of toxic sentences that are targeted at non-marginalized groups. The results indicate that the selection bias is the highest in the sensitive attribute of religion (0.077), followed by race (0.053), and finally gender (0.027).

$$Selection_{g,\hat{g}} = \left| \left(\frac{N_{g,toxicity=1}}{N_g} \right) - \left(\frac{N_{\hat{g},toxicity=1}}{N_{\hat{g}}} \right) \right| \tag{4}$$

To measure the impact of selection bias on the fairness of the toxicity detection task, we used the Pearson's correlation coefficient (ρ) between the fairness scores measured by the different fairness metrics and selection bias scores in the civil community training dataset. We found that, for AlBERT-base, selection bias scores have a strong positive correlation with the fairness scores when measured as FPR_gap (ρ = 0.984), AUC_gap (ρ = 0.911), and TPR_gap (ρ = 0.633). Similar correlation patterns were found in RoBERTa-base. As for BERT-base, there are weaker positive correlations with TPR_gap and AUC_gap and almost no correlation with FPR_gap.

These results suggest that selection bias in the training datasets used for downstream tasks has a direct impact on the fairness of the examined language models, as evidenced by the positive

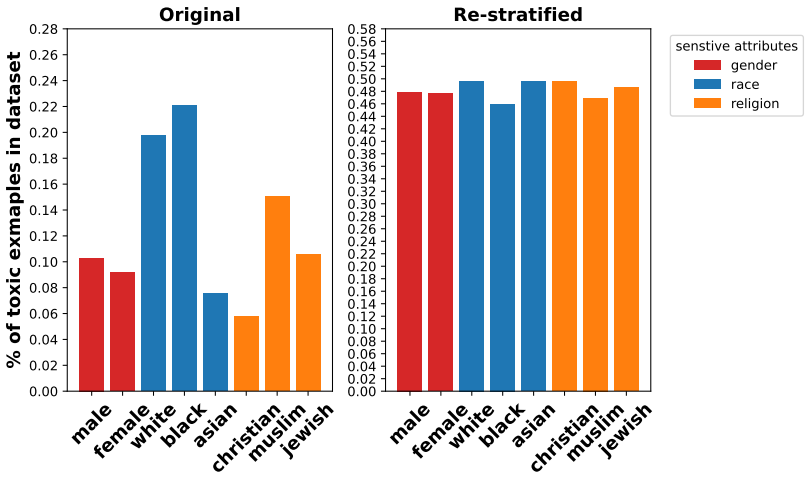

Figure 3: The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after re-stratification (right).

correlations with the fairness of these models on the downstream task of toxicity detection, as measured by the different fairness metrics.

6.4 Selection bias removal

According to Shah, Schwartz, and Hovy (2020), to remove selection bias, a realignment in the sample distribution in the training dataset is required to minimize the mismatch in the class representation between the different identities. The authors in Shah, Schwartz, and Hovy (2020) list data re-stratification as a mitigation technique to remove selection bias and to match the ideal distribution of balanced class representations for the different identities. We followed this suggestion to have a balanced representation of positive and negative examples for the different marginalized and non-marginalized groups that we study in this section.

Following the same methodology used in Zmigrod et al. (2019), data augmentation was used to create slightly altered examples to balance the class representations and add them to the civil community training dataset. Since the percentages of the positive examples for the different identity groups are small, ranging from 0.05 to 0.2, as shown in Figure 3The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after re-stratification (right).figure.caption.3 (Left), we create more positive examples by slightly altering the existing positive examples, without changing the meaning, in the dataset using word substitutions. Word substitutions were generated using the NLPAUG tool[1] that uses contextual word embeddings to find the word substitutions (Ma 2019). After adding the synthesized positive examples to the civil community training dataset, the new re-stratified civil community training dataset contains 443,046 data items with balanced class representation, as shown in Figure 3The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after re-stratification (right).figure.caption.3 (Right). The selection bias in the re-stratified civil community training

[1] https://github.com/makcedward/nlpaug

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	**0.847**	0.006	0.039	0.004
	+ downstream-stratified-data	0.816	↓ 0.005	↓ 0.003	↑ 0.005
	BERT	**0.830**	0.090	0.036	0.010
	+ downstream-stratified-data	0.817	↓ 0.007	↓ 0.006	↓ 0.006
	RoBERTa	**0.851**	0.005	0.032	0.011
	+ downstream-stratified-data	0.842	↑ 0.006	↓ 0.005	↓ 0.002
Race	ALBERT	**0.847**	0.008	0.002	0.019
	+ downstream-stratified-data	0.816	↑ 0.022	↑ 0.026	↓ 0.008
	BERT	**0.830**	0.016	0.002	0.026
	+ downstream-stratified-data	0.817	↓ 0.010	↑ 0.018	↓ 0.008
	RoBERTa	**0.851**	0.003	0.011	0.021
	+ downstream-stratified-data	0.842	↑ .014	0.011	↓ 0.014
Religion	ALBERT	**0.847**	0.010	0.109	0.020
	+ downstream-stratified-data	0.816	↑ 0.030	↓ 0.058	↓ 0
	BERT	**0.830**	0.008	0.063	0.012
	+ downstream-stratified-data	0.817	↑ 0.020	↓ 0.049	↓ 0.006
	RoBERTa	**0.851**	0.022	0.160	0.027
	+ downstream-stratified-data	0.842	↓ 0.019	↓ 0.071	↓ 0.001

Table 6: Toxicity detection performance and fairness scores for all models before and after removing selection bias. **Bold** values refer to higher AUC scores and better performance. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased and the fairness improved. The word *downstream* is used to explain that the bias removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.

dataset is reduced to 0.002, 0.019 and 0.017 for the gender, race, and religion sensitive attributes respectively.

Then, the examined AlBERT, BERT, and RoBERTa models were fine-tuned on the new re-stratified civil community training dataset. Balancing the class representation in the dataset led to a reduction in the performance (AUC scores) of all three models (+ downstream-stratified-data), as shown in Table 6Toxicity detection performance and fairness scores for all models before and after removing selection bias. **Bold** values refer to higher AUC scores and better performance. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased and the fairness improved. The word *downstream* is used to explain that the bias removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.table.0.6. This reduction in the AUC scores is a result of predicting more positive examples than the original models and in turn resulting in more false positive and more less true negative.

To investigate the impact of selection bias removal on the fairness of the task of toxicity detection, we measure the fairness in all three examined models after fine-tuning them on the new re-stratified civil community dataset. The fairness scores were analyzed for the examined sensitive attributes using the different fairness metrics for all the examined models. Results in Table 6Toxicity detection performance and fairness scores for all models before and after removing selection bias. **Bold** values refer to higher AUC scores and better performance. (↑) denotes that the fairness metric score increased and the fairness worsened. (↓) denotes that the fairness metric score decreased and the fairness improved. The word *downstream* is used to explain that the bias

removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.table.0.6 showed that for the AUC_gap metric, the fairness improved for all models and most of the sensitive attributes as evidenced by AlBERT (race, religion), BERT (gender, race, religion), and RoBERTa (gender, race, religion). However, the results are inconsistent for the TPR_gap or FPR_gap across models or sensitive attributes. We speculate that TPR_gap and FPR_gap do not reflect the improvement in fairness as measured using the AUC_gap metric because after removing selection bias, the models tend to predict more positives (false positive and true positive). The FPR and TPR increased for some identity groups (White and Black) and got reduced for other groups (Asian), which made the TPR_gap and FPR_gap increase. On the other hand, this does not happen with the AUC scores for the different identity groups, hence the AUC_gap scores did not increase, which might be the case because the AUC scores and the AUC_gap are threshold-agnostic metrics and don't directly rely on the models' false or true positive predictions.

When examining the cases where the fairness improved according to all the fairness metrics, we found only two cases out of nine. The first is when BERT is fine-tuned on the re-stratified training dataset, which led to improvement of the model's fairness regarding the gender sensitive attribute. The second is fine-tuning RoBERTa on the re-stratified training dataset, which led to improvement of the model's fairness regarding the religion sensitive attribute.

In summary, it was shown that selection bias is influential on the models' fairness on the task of toxicity detection and removing it by balancing the class representations for all the identity groups in the training dataset using data augmentation improved the models' fairness according to the AUC_gap metric, but not according to all the examined fairness metrics.

6.5 Overamplification bias

According to Shah, Schwartz, and Hovy (2020), overamplification bias happens during LM training as LM models rely on small differences between sensitive attributes regarding an objective function and amplify these differences to be more pronounced in the predicted outcome. For the task of toxicity detection, overamplification bias could happen because certain identity groups exist more often within specific semantic contexts in the training datasets. For example, when an identity name, e.g., "Muslim" co-exists in the same sentence with the word "terrorism" more often than other identity names, e.g., "Buddist". Even if the sentence does not contain any hate speech, e.g., "*Anyone could be a terrorist, not just muslims*", the LM model will learn to pick this information up and amplify it, leading to the fine-tuned LM model predicting future sentences that contain the word "Muslim" as hateful.

In Zhao et al. (2017), the authors propose a method to measure and mitigate overamplification bias when training models on biased corpora. The authors propose the RBA framework for reducing bias amplification in predictions. Their proposed method introduces corpus-level constraints so that gender indicators co-occur no more often together with elements of the prediction task than in the original training distribution (Zhao et al. 2017). The aim of the proposed method is to limit the model bias to the bias in the dataset without amplification. This method would only be effective if the training dataset is not biased. So in this section, we aim to examine overamplification bias in the training dataset before it gets amplified during model training.

$$Overamplification_{g,\hat{g}} = |N_g - N_{\hat{g}}| \tag{5}$$

Equation 5Overamplification biasequation.0.6.5 was used to measure overamplification bias in the civil community training dataset. To this end, we measured the differences between the number

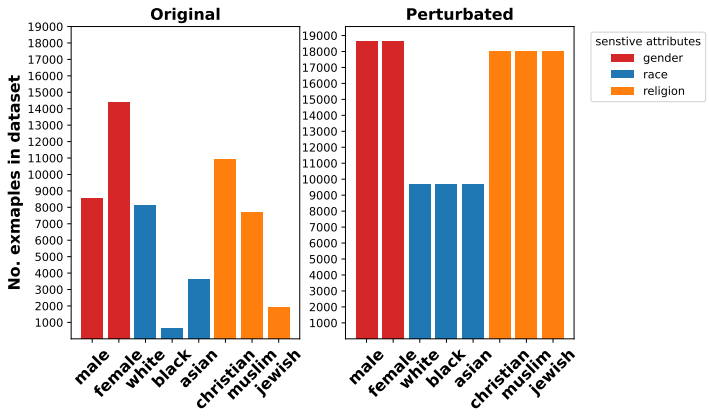

Figure 4: The number of examples, for each identity group in the civil community training dataset in the original dataset (left) and after perturbation (right).

of examples targeted at marginalized vs. non-marginalized groups, as shown in Figure 4The number of examples, for each identity group in the civil community training dataset in the original dataset (left) and after perturbation (right).figure.caption.4 (Left). Then the scores are normalized using the Max normalization (li and Liu 2011) where each value in $overamplification_{g,\hat{g}}$ for gender (5795), race(5795), and religion (6118.5) is divided by max value which is 6118.5. The reason behind using Max normalization is to avoid having a score of 0 which might be misleading in the context of bias. The different sizes mean that certain identity groups appear in more semantic contexts than others. These contexts could be positive or negative. The overamplification bias scores in the civil community training dataset for the sensitive attributes are 1.0 for religion, followed by 0.97 for race, and finally 0.94 for gender.

To investigate the impact of overamplification bias on the models' fairness on the downstream task of toxicity detection, we measured the Pearson correlation coefficient (ρ) between overamplification bias scores and the fairness scores measured using threshold-based and threshold-agnostic fairness metrics. A strong positive correlation was found for AlBERT-base between overamplification bias and fairness scores, as measured by FPR_gap (ρ = 0.988), AUC_gap (ρ = 0.921) and TPR_gap (ρ = 0.613). A similar pattern of correlations was found for RoBERTa-base. As for BERT-base, there are weaker positive correlations with TPR_gap and AUC_gap and almost no correlation with FPR_gap. These results suggest that overamplification bias in the civil community training dataset measured as the difference in the sentences that are targeted at the different groups might have a direct impact on the models' fairness, and that during fine-tuning these differences are amplified in a way that might then make a bigger impact on the models' fairness. Additionally, overamplification bias could also mean the amplification of selection bias introduced earlier in section 6.3Selection biassubsection.0.6.3.

6.6 Overamplification bias removal

To overcome overamplification bias, we followed the work of Webster et al. (2020) where the authors propose to train the model on a training dataset with balanced semantic representations of the different identity groups using counterfactuals. To achieve that balance in the civil community training dataset, each identity, marginalized and non-marginalized, should be presented in similar

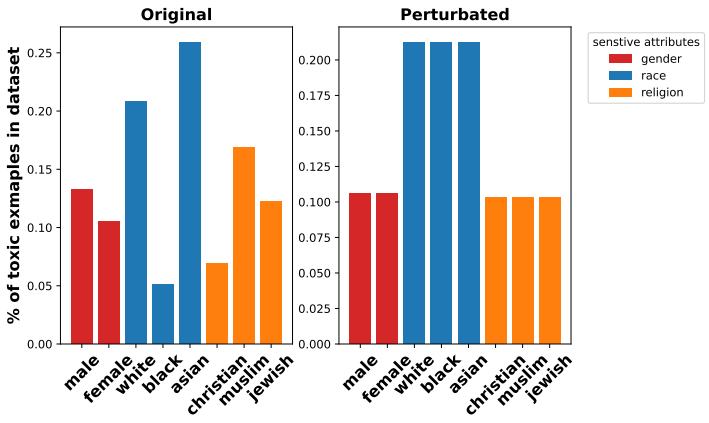

Figure 5: The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after perturbation (right).

semantic contexts so that the models would not associate certain semantic contexts with certain identity groups.

To create the perturbations, the first attempt was to fine-tune a Text-to-Text model (Raffel et al. 2020) on the PANDA dataset (Qian et al. 2022) to automatically generate perturbations. We used the same values for the hyperparameters as shared in Raffel et al. (2020) and the Text-to-Text model achieved a ROUGE-2 score of 0.9, which is similar to the score reported in the original study (ROUGE-2=0.9) (Qian et al. 2022). However, upon inspection of the perturbed text, it was found that the perturbed text is not consistently changing and that it does not perform well on religious or racial identities. Upon further inspection, we realized that the perturbed text in the PANDA dataset is inconsistent, and sometimes the perturbations are not correct. This is not picked up by the ROUGE-2 metric because it works by comparing the overlap of bi-grams between two sentences (original and perturbed) (Ng and Abrecht 2015), which is not indicative of good performance in the task of perturbation generation in comparison to a task like text translation. Since the original and the perturbed sentences are similar except for words that describe identity groups, the ROUGE-2 metric gives high scores regardless of the quality of the generated perturbation.

To address this issue, we followed the same method used to create the balanced civil community fairness dataset, as explained in section 5Fairness in the task of toxicity detectionsection.0.5. To this end, perturbations (counterfactuals) were created for each sentence in the civil community training dataset. As discussed in section 5.1Balanced fairness datasetsubsection.0.5.1, the most common nouns and adjectives in the civil community dataset suggest that there is no risk of *Asymmetric Counterfactuals* which means we can create the perturbations for toxic and non-toxic sentences. So for the identity of Black people, in addition to the subset of sentences that are targeted at Black people, we created perturbations from the sentences that are targeted at White and Asian identities, replacing any references to White or Asian identity names with Black identity names. The same was done with the White identity. In addition to the sentences that are targeted at the White identity, perturbations were created from the sentences that are targeted at Black and Asian people, replacing the words that describe Asian or Black identities with identity words that describe the White identity. The same process was repeated for the Asian identity. In addition to the sentences that target Asian people, perturbations were created from the sentences that are targeted at Black and White people, replacing any reference to Black and White with identity

words that describe Asian people. This way, it was ensured that all the different racial identities in the dataset are represented in the same way. The same process was repeated for the identity groups in the gender and religion sensitive attributes. The new distribution of identities is shown in Figure 4The number of examples, for each identity group in the civil community training dataset in the original dataset (left) and after perturbation (right).figure.caption.4 (Right). Figure 5The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after perturbation (right).figure.caption.5 shows how removing overamplification bias mitigated selection bias, as shown in Figure 5The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after perturbation (right).figure.caption.5 (Left), by balancing the ratios of positive examples for the different identity groups in the same sensitive attribute, as shown in Figure 5The percentage of positive (toxic) examples, for each identity group in the civil community training dataset in the original dataset (left) and after perturbation (right).figure.caption.5 (Right).

The size of the balanced-perturbed civil community training dataset after generating the perturbations is 382,212 sentences and the ratio between the positive and the negative examples for each identity group within the same sensitive attribute is the same. For example, in the gender attribute, the ratio of the positive (toxic) examples in the male and female identity groups is 0.10, in the race attribute, the ratio of the positive examples for the Black, White and Asian groups is 0.2, and for the religion attribute, the ratio of the positive examples for the Muslim, Christian and Jewish groups is 0.10. It is worth noting that with similar ratios of positive examples between the different identity groups in the same sensitive attribute, selection bias in the new civil community training dataset is mitigated as well. Finally, we used the balanced perturbed civil community training dataset to fine-tune the examined models (+ downstream-perturbed-data).

Since selection bias could also be amplified by the models during training, we perturbed the re-stratified civil community training dataset, as explained in section 6.3Selection biassubsection.0.6.3. This does not only guarantee balanced semantic contexts, but also a more balanced ratio between positive and negative examples in the civil community training dataset. The new perturbed-stratified dataset contains 841,814 sentences, where the ratio of the positive examples for the male and female identity groups is 0.48, the ratio of positive examples between Black, Asian, and White identity groups is 0.48, and the ratio of positive examples between the Muslim, Christian, and Jewish identity groups is 0.49.

As an alternative to using counterfactual and data perturbation to remove overamplification bias, we used SentDebias to remove the biased subspaces from the models as explained in section 6.2Representation bias removalsubsection.0.6.2. To this end , to remove overamplification bias, we remove the biased subspaces after fine-tuning the models on the civil community dataset (+ downstream-sentDebias).

To investigate the impact of removing overamplification bias on the fairness of the toxicity detention task, we used the threshold-based and threshold-agnostic fairness metrics to measure the impact of the examined debiasing techniques for removing overamplification bias on the models' fairness.

- **Downstream-SentDebias:** Starting with the impact of removing the biased subspaces from the fine-tuned models, it is shown that the performance of the models after removing the biased representations is much worse, almost random, with the AUC scores close to 0.5 as shown in Table 7 (+ downstream-sentDebias). This is expected since the models lost a lot of information related to toxicity along with the biased subspaces. To simplify the result's analysis, we examined the impact of removing a certain type of bias on the fairness of the corresponding sensitive attribute, as explained in section 6.1Representation biassubsection.0.6.1.

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	0.847	0.006	0.039	0.004
	+ downstream-sentDebias-gender	0.524	↓ 0	↓ 0.008	↑ 0.011
	+ downstream-perturbed-data	**0.848**	↓ 0.001	↓ 0.010	0.004
	+ downstream-perturbed-stratified-data	0.803	↓ 0.005	↓ 0.006	↑ 0.008
	BERT	0.830	0.09	0.036	0.01
	+ downstream-sentDebias-gender	0.478	↓ 0	↓ 0.001	↓ 0.004
	+ downstream-perturbed-data	**0.837**	↓ 0.003	↓ 0.005	↓ 0.003
	+ downstream-perturbed-stratified-data	0.810	↓ 0.003	↓ 0.003	↓ 0.005
	RoBERTa	0.851	0.005	0.032	0.011
	+ downstream-sentDebias-gender	0.520	↑ 0.015	↓ 0.019	↓ 0.004
	+ downstream-perturbed-data	**0.873**	↓ 0.001	↓ 0.009	↓ 0.002
	+ downstream-perturbed-stratified-data	0.825	↓ 0	↓ 0.005	↓ 0.007
Race	ALBERT	0.847	0.008	0.002	0.019
	+ downstream-sentDebias-race	0.421	↓ 0	↑ 0.004	↓ 0.001
	+ downstream-perturbed-data	**0.848**	↓ 0.003	↓ 0.001	↓ 0.003
	+ downstream-perturbed-stratified-data	0.803	↑ 0.004	0.002	↓ 0.002
	BERT	0.830	0.016	0.002	0.026
	+ downstream-sentDebias-race	0.504	↓ 0	↓ 0	↓ 0.002
	+ downstream-perturbed-data	**0.837**	↓ 0.009	↑ 0.019	↓ 0.003
	+ downstream-perturbed-stratified-data	0.810	↓ 0.002	0.002	↓ 0.002
	RoBERTa	0.851	0.003	0.011	0.021
	+ downstream-sentDebias-race	0.561	↓ 0	↓ 0	↓ 0.005
	+ downstream-perturbed-data	**0.873**	↑ 0.018	↑ 0.038	↓ 0.003
	+ downstream-perturbed-stratified-data	0.825	0.003	↓ 0.006	↓ 0.001
Religion	ALBERT	0.847	0.010	0.109	0.020
	+ downstream-sentDebias-religion	0.507	↓ 0.004	↓ 0	↓ 0.002
	+ downstream-perturbed-data	**0.848**	↓ 0.002	↓ 0.011	↓ 0.001
	+ downstream-perturbed-stratified-data	0.803	↓ 0	↓ 0.002	↓ 0.002
	BERT	0.830	0.008	0.063	0.012
	+ downstream-sentDebias-religion	0.447	↓ 0	↓ 0	↑ 0.030
	+ downstream-perturbed-data	**0.837**	↓ 0.002	↓ 0.011	↓ 0.001
	+ downstream-perturbed-stratified-data	0.810	↓ 0	↓ 0.001	↓ 0.003
	RoBERTa	0.851	0.022	0.160	0.027
	+ downstream-sentDebias-religion	0.523	↓ 0	↓ 0	↓ 0
	+ downstream-perturbed-data	**0.873**	↓ 0.001	↓ 0.003	↓ 0.002
	+ downstream-perturbed-stratified-data	0.825	↓ 0.001	↓ 0.004	↓ 0.001

Table 7: Toxicity detection performance and fairness scores for all models before and after overamplification bias removal. **Bold** values refer to higher AUC scores and better performance . (↑) denotes that the fairness score increased and the fairness worsened. (↓) denotes that the fairness score decreased and the fairness improved.The word *downstream* is used to explain that the bias removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.

The results show that removing the biased subspaces after fine-tuning the models led to improved fairness in all the models according to all fairness metrics for almost all the sensitive attributes. However, these results are misleading. After analyzing the prediction probabilities of the models after removing the biased subspaces, we found that in most of the models, the number of positive predictions is either immensely reduced or immensely increased. This results in very low false positive rates and very low true positive rates, or very high false positive rates and very high true positive rates, resulting in minimal TPR_gap and FPR_gap scores (≈ 0). These results suggest that removing the biased subspace from fine-tuned models to mitigate overamplification bias falsely improves the models' fairness while it deteriorates their performance.

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
	ALBERT	0.847	0.006	0.039	0.004
	+ upstream-sentDebias-gender	0.840	0.006	↓ 0.032	0.004
	+ downstream-sentDebias-gender	0.524	↓ 0	↓ 0.008	↑ 0.011
	+ downstream-perturbed-data	**0.848**	↓ 0.001	↓ 0.01	0.004
	+ downstream-stratified-data	0.816	↓ 0.005	↓ 0.003	↑ 0.005
	+ downstream-perturbed-stratified-data	0.803	↓ 0.005	↓ 0.006	↑ 0.008
	+ upstream-sentDebias-gender- downstream-all-data-debias	0.792	↓ 0.001	↓ 0.005	↑ 0.008
	BERT	0.83	0.09	0.036	0.01
	+ upstream-sentDebias-gender	**0.841**	↓ 0.011	↑ 0.049	↓ 0.006
	+ downstream-sentDebias-gender	0.478	↓ 0	↓ 0.001	↓ 0.004
Gender	+ downstream-perturbed-data	0.837	↓ 0.003	↓ 0.005	↓ 0.003
	+ downstream-stratified-data	0.817	↓ 0.007	↓ 0.006	↓ 0.006
	+ downstream-perturbed-stratified-data	0.810	↓ 0.003	↓ 0.003	↓ 0.005
	+ upstream-sentDebias-gender- downstream-all-data-debias	0.791	↓ 0	↓ 0.003	↓ 0.002
	RoBERTa	0.851	0.005	0.032	0.011
	+ upstream-sentDebias-gender	0.856	↑ 0.006	↓ 0.022	↓ 0.003
	+ downstream-sentDebias-gender	0.520	↑ 0.015	↓ 0.019	↓ 0.004
	+ downstream-perturbed-data	**0.873**	↓ 0.001	↓ 0.009	↓ 0.002
	+ downstream-stratified-data	0.842	↑ 0.006	↓ 0.005	↓ 0.002
	+ downstream-perturbed-stratified-data	0.825	↓ 0	↓ 0.005	↓ 0.007
	+ upstream-sentDebias-gender-downstream-all-data-debias	0.837	↓ 0.001	↓ 0.003	↓ 0.004
	ALBERT	0.847	0.008	0.002	0.019
	+ upstream-sentDebias-race	0.838	↓ 0.003	↑ 0.003	↓ 0.013
	+ downstream-sentDebias-race	0.421	↓ 0	↑ 0.004	↓ 0.001
	+ downstream-perturbed-data	**0.848**	↓ 0.003	↓ 0.001	↓ 0.003
	+ downstream-stratified-data	0.816	↑ 0.022	↑ 0.026	↓ 0.008
	+ downstream-perturbed-stratified-data	0.803	↑ 0.004	0.002	↓ 0.002
	+ upstream-sentDebias-race-downstream-all-data-debias	0.817	↓ 0.001	↑ 0.017	↓ 0.001
	BERT	0.830	0.016	0.002	0.026
	+ upstream-sentDebias-race	0.829	↑ 0.021	↑ 0.005	↓ 0.024
	+ downstream-sentDebias-race	0.504	↓ 0	↓ 0	↓ 0.002
Race	+ downstream-perturbed-data	**0.837**	↓ 0.009	↑ 0.019	↓ 0.003
	+ downstream-stratified-data	0.817	↓ 0.010	↑ 0.018	↓ 0.008
	+ downstream-perturbed-stratified-data	0.810	↓ 0.002	0.002	↓ 0.002
	+ upstream-sentDebias-race-downstream-all-data-debias	0.815	↓ 0.005	0.002	↓ 0
	RoBERTa	0.851	0.003	0.011	0.021
	+ upstream-sentDebias-race	0.854	↑ 0.017	↓ 0.009	0.021
	+ downstream-sentDebias-race	0.561	↓ 0	↓ 0	↓ 0.005
	+ downstream-perturbed-data	**0.873**	↑ 0.018	↑ 0.038	↓ 0.003
	+ downstream-stratified-data	0.842	↑ .014	0.011	↓ 0.014
	+ downstream-perturbed-stratified-data	0.825	0.003	↓ 0.006	↓ 0.001
	+ upstream-sentDebias-race-downstream-all-data-debias	0.842	↑ 0.006	↑ 0.014	↓ 0.003
	ALBERT	0.847	0.010	0.109	0.020
	+ upstream-sentDebias-religion	0.837	↑ 0.019	↓ 0.094	↓ 0.016
	+ downstream-sentDebias-religion	0.507	↓ 0.004	↓ 0	↓ 0.002
	+ downstream-perturbed-data	**0.848**	↓ 0.002	↓ 0.011	↓ 0.001
	+ downstream-stratified-data	0.816	↑ 0.03	↓ 0.058	↓ 0
	+ downstream-perturbed-stratified-data	0.803	↓ 0	↓ 0.002	↓ 0.002
	+ upstream-sentDebias-religion-downstream-all-data-debias	0.811	↓ 0.001	↓ 0.013	↓ 0.003
	BERT	0.830	0.008	0.063	0.012
	+ upstream-sentDebias-religion	0.833	↑ .015	↑ 0.084	↑ 0.017
	+ downstream-sentDebias-religion	0.447	↓ 0	↓ 0	↑ 0.03
Religion	+ downstream-perturbed-data	**0.837**	↓ 0.002	↓ 0.011	↓ 0.001
	+ downstream-stratified-data	0.817	↑ 0.02	↓ 0.049	↓ 0.006
	+ downstream-perturbed-stratified-data	0.810	↓ 0	↓ 0.001	↓ 0.003
	+ upstream-sentDebias-religion-downstream-all-data-debias	0.820	↓ 0	↓ 0.005	↓ 0.003
	RoBERTa	0.851	0.022	0.16	0.027
	+ upstream-sentDebias-religion	0.843	↓ 0.021	↓ 0.10	↓ 0.003
	+ downstream-sentDebias-religion	0.523	↓ 0	↓ 0	↓ 0
	+ downstream-perturbed-data	**0.873**	↓ 0.001	↓ 0.003	↓ 0.002
	+ downstream-stratified-data	0.842	↓ 0.019	↓ 0.071	↓ 0.001
	+ downstream-perturbed-stratified-data	0.825	↓ 0.001	↓ 0.004	↓ 0.001
	+ upstream-sentDebias-religion-downstream-all-data-debias	0.834	↓ 0	↓ 0.006	↓ 0.007

Table 8: Summary of the performance and fairness scores for all models before and after applying different bias removal methods to remove different sources of bias. **Bold** values refer to higher AUC scores. (↑) denotes that the fairness score increased and the fairness worsened. (↓) denotes that the fairness score decreased and the fairness improved. The word *upstream* is used here to refer to removing bias from the models before fine-tuning them on the task of toxicity detection. The word *downstream* is used to explain that the bias removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.

- **Data perturbation:** As for the impact of fine-tuning the models on perturbed datasets, the results in Table 7 (+ downstream-perturbed-data) show that the performance slightly improved for all the models. Fine-tuning the models on the perturbed data made the models predict more positives, true positives and false positives, without hurting the true negatives much, which is the case with the other examined debiasing methods.
  An examination of the fairness scores after fine-tuning the models of the perturbed dataset shows that the different fairness metrics agree, in almost all the models for most of the sensitive attributes, that fine-tuning the models on the perturbed dataset improves the fairness. These results also suggest that mitigating overamplification bias by fine-tuning language models on perturbed datasets with balanced semantic contexts for different identity groups improved the fairness and the performance of the examined language models.

- **Perturbed-re-stratified data:** When the models were fine-tuned on the perturbed-re-stratified civil community dataset, the performance was slightly worse for all three models, as shown in Table 7Toxicity detection performance and fairness scores for all models before and after overamplification bias removal. **Bold** values refer to higher AUC scores and better performance . (↑) denotes that the fairness score increased and the fairness worsened. (↓) denotes that the fairness score decreased and the fairness improved.The word *downstream* is used to explain that the bias removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.table.0.7 (+ downstream-perturbed-stratified-data). Like most of the other debiasing techniques, fine-tuning the perturbed re-stratified data caused the model to predict more positives, but especially more false positives in AlBERT and RoBERTa.
  An examination of the fairness scores shows that the AUC_gap consistently improved across all models and for almost all the sensitive attributes, AlBERT (race, religion), BERT (gender, race, religion), and RoBERTa (gender, race, religion). The results for FPR_gap and TPR_gap are not as consistent, but still improved for most of the sensitive attributes and models. These improvements are better than removing only selection bias by fine-tuning the models on re-stratified data. Yet not as good as fine-tuning the models on only perturbed-data, which improved the models' fairness more consistently.

6.7 Multibiases

Since all the examined sources of bias do not happen isolated from one another, we examined the impact of implementing all the proposed methods to remove bias from the different sources, as explained in sections 6.2Representation bias removalsubsection.0.6.2, 6.4Selection bias removalsubsection.0.6.4, and 6.6Overamplification bias removalsubsection.0.6.6. To this end, we fine-tuned the models after removing the representation bias (upstream) using SentDebias on the perturbed-re-stratified civil community training dataset to remove both selection and overamplification biases (downstream). The AUC scores were slightly worse compared to the original models, as shown in Table 8Summary of the performance and fairness scores for all models before and after applying different bias removal methods to remove different sources of bias. **Bold** values refer to higher AUC scores. (↑) denotes that the fairness score increased and the fairness worsened. (↓) denotes that the fairness score decreased and the fairness improved. The word *upstream* is used here to refer to removing bias from the models before fine-tuning them on the task of toxicity detection. The word *downstream* is used to explain that the bias removal technique is applied during fine-tuning the model on the downstream task of toxicity detection.table.0.8 (+upstream-sentDebias-gender-downstream-all-data-debias).

The results indicate that, similar to previous results, removing different sources of bias made the model predict more positives (false positives and true positives). Fore exmaple, removing

gender bias in ALBERT (+ upstream-sentDebias-gender-downstream-all-data-debias), BERT (+ upstream-sentDebias-gender-downstream-all-data-debias), and RoBERTa (+ upstream-sentDebias-gender-downstream-all-data-debias) led to worse AUC scores because the number of false negatives also increased, which is similar to removing racial and religious bias. The fairness scores show improvement across most of the inspected models, fairness metrics, and sensitive attributes. The results show that removing different sources of bias, in most cases, improved the fairness scores ways similar to the results of fine-tuning the models on perturbed-re-stratified datasets. For example, the pattern of fairness improvement scores is the same in Gender (AlBERT, BERT, RoBERTa), Race (BERT), Religion (ALBET, BERT, RoBERTa).

On the other hand, there are barely any similarities to the fairness improvement pattern between removing all sources of bias and removing representation bias. These results suggest that removing the downstream sources of bias (selection and overamplification) has a stronger impact on the models' fairness than the upstream sources of bias (representation bias). A similar finding was made by Steed et al. (2022).

7. Discussion

After investigating the impact of the different sources of bias and their removal on the downstream task of toxicity detection, we further analyse and summarize our results to answer our research questions.

7.1 How impactful are the different sources of bias on the fairness of language models on the downstream task of toxicity detection?

The results of the previous sections show that there is a positive correlation between representation bias, selection bias, and overamplification bias and the fairness scores measured by the different fairness metrics. This suggests that all the inspected sources of bias have an impact on the fairness of the inspected language models on the downstream task of toxicity detection. To find out the most impactful source of bias, we compared the strength of the correlation between the different sources of bias and the models' fairness. The CrowS-Pairs scores were used to measure representation bias, as CrowS-Pairs is the only metric that correlates positively with all the different fairness metrics, as explained in section 6.1Representation biassubsection.0.6.1. For selection and overamplification sources of bias, we used the scores reported in sections 6.3Selection biassubsection.0.6.3 and 6.5Overamplification biassubsection.0.6.5 respectively.

The results shown in Table 9The Pearson correlation coefficient (ρ) between different sources of bias and fairness metrics in all the models. **Bold** values refer to the highest (ρ) values and hence the strongest correlation.table.0.9 indicate that correlation coefficients are high for both selection and overamplification bias in both AlBERT and BERT models. As for the RoBERTa model, representation bias has the highest correlation, which could be the case because RoBERTa is the most representation-biased model for all the sensitive attributes (gender, race, religion) when measured using the Crows-pairs metric, as shown in Table 4Representation bias scores in the examined models using different bias metrics before and after removing bias using the SentDebias algorithm. ($\uparrow$) denotes that the fairness metric score increased and the fairness worsened. ($\downarrow$) denotes that the fairness metric score decreased, and the fairness improved.table.0.4. To summarize these findings and answer the research question, the results suggest that downstream bias sources of bias (selection bias and overamplification bias) is the most influential bias in comparison to upstream bias (representation bias). The results also suggest that overamplification bias is more impactful, as evidenced by the higher correlation coefficients, than representation and selection sources of bias. To have more conclusive answers, we investigated the most effective debiasing method as an indicator of the most impactful source of bias.

AlBERT			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	0.466	**0.999**	0.233
Selection	0.984	0.633	0.911
Overampflication	**0.988**	0.613	**0.921**
BERT			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	-0.536	**0.819**	-0.369
Selection	-0.037	0.418	0.150
Overampflication	**-0.011**	0.395	**0.175**
RoBERTa			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	**0.972**	**0.980**	0.555
Selection	0.809	0.785	0.992
Overampflication	0.794	0.770	**0.995**

Table 9: The Pearson correlation coefficient (ρ) between different sources of bias and fairness metrics in all the models. **Bold** values refer to the highest (ρ) values and hence the strongest correlation.

7.2 What is the impact of removing the different sources of bias on the fairness of the downstream task of toxicity detection?

To answer this research question, we summarized the findings on the impact of removing the different sources of bias on the models' fairness, as discussed in sections 6.2Representation bias removalsubsection.0.6.2, 6.4Selection bias removalsubsection.0.6.4, 6.6Overamplification bias removalsubsection.0.6.6, and 6.7Multibiasessubsection.0.6.7. The results in Table 10Summary of the most effective debiasing method according to all the fairness metrics for all the models and all the sensitive attributes.table.0.10 show that the most effective debiasing method that improved fairness according to all the used fairness metrics in most of the models and sensitive attributes is removing overamplification bias. These results support the finding from the previous research question that the most impactful source of bias is overamplification and removing it is the most effective on the models' fairness. The results also show that fine-tuning language models on perturbed data with balanced contextual semantic representation is more effective than training on perturbed re-stratified data. Furthermore, fine-tuning the models on perturbed data also addresses selection bias, as the ratio of positive examples is the same for all identity groups in the same sensitive attributes. It is also important to mention that removing downstream sources of bias (selection bias and overamplification bias) is more effective than removing upstream representation bias. This finding was also made by Kaneko, Bollegala, and Okazaki (2022); Steed et al. (2022). Removing the biased subspaces after fine-tuning (Downstream-SentDebias) is effective in some cases, like AlBERT (religion), BERT (gender, race), RoBERTa (race, religion). However, using this technique leads to poor performance. Consequently, we do not recommend using this debiasing technique, as it is important to find the right trade-off between performance and fairness. Removing selection bias by fine-tuning the models on re-stratified data improved

	AlBERT-base			BERT-base			RoBERTa-base		
Debias approach	gender	race	religion	gender	race	religion	gender	race	religion
Upstream-SentDebias	✗	✗	✗	✗	✗	✗	✗	✗	✓
Downstream-SentDebias	✗	✗	✓	✓	✓	✗	✗	✓	✓
Downstream-perturbed-data	✗	✓	✓	✓	✗	✓	✓	✗	✓
Downstream-stratified-data	✗	✗	✗	✓	✗	✗	✗	✗	✓
Downstream-perturbed-stratified-data	✗	✗	✓	✓	✗	✓	✓	✗	✓
Upstream-sentDebias-Downstream-all-data-debias	✗	✗	✓	✓	✗	✓	✓	✗	✓

Table 10: Summary of the most effective debiasing method according to all the fairness metrics for all the models and all the sensitive attributes.

fairness in some cases, like BERT (gender) and RoBERTa (religion), but was not as effective as removing overamplification bias. We speculate that this is the case because removing selection bias includes re-stratifying the data by adding synthesized positive examples to the training dataset, leading to a balanced class ratio between positive and negative examples (≈ 0.5) for all identity groups. This resulted in the model predicting more false positives and less true negatives, which resulted in worse fairness and worse performance. On the other hand, training the model on perturbed data ensured balanced positive class representation between the different identity groups, but the ratio between the positive and negative class stayed low (≈ 0.1 to 0.2). This made the model predict more positives, specifically more true positives, without hurting the number of true negatives, which improved both the performance and fairness of the inspected language models.

Removing both selection and overamplification bias (Downstream-perturbed-stratified-data), was more effective than removing only selection bias and less effective than removing only overamplification bias. Removing all sources of bias (upstream and downstream) resulted in the same pattern as fine-tuning the model on re-stratified perturbed data (+ downstream-perturbed-stratified-data), which confirms that removing upstream bias does not have a strong impact on the models' fairness. However, results showed that in rare cases, removing both upstream and downstream bias leads to improved fairness. For example, the FPR_gap score for the gender-sensitive attribute of the AlBERT model (+downstream-perturbed-stratified-data) is 0.005, while the FPR_gap score for the gender-sensitive attribute of the AlBERT model (+ upstream-sentDebias-gender-downstream-all-data-debias) is 0.001. However, this was not always the case.

7.3 Which debiasing techniques to use to ensure the models' fairness on the task of toxicity detection?

The answer to the previous research questions suggests that the best technique to ensure the fairness of the task of toxicity detection is to remove overamplification bias by fine-tuning the models on a perturbed dataset. To confirm that this is indeed the most effective downstream debiasing technique and to answer this research question, we used the counterfactual fairness metric (Kusner et al. 2017) to get a more profound understanding of the effect of the different bias removal techniques on the models' fairness. To this end, we used the perturbation sensitivity score ($SenseScore$) metric proposed by Prabhakaran, Hutchinson, and Mitchell (2019) to inspect closely the models' prediction probabilities of the different identity groups within each sensitive attribute. $SenseScore$ is defined in Equation 6Which debiasing techniques to use to ensure the models' fairness on the task of toxicity detection?equation.0.7.6 as the absolute mean of the difference between the model prediction ($f(x)$) of the factual (x) and the counterfactual (perturbed) ($\hat{x}$) instances, divided by the number of examples (X). For a model $f(x) : X \rightarrow Y$, a factual instance (x) that contains identity (g), and counterfactual instance ($\hat{x}$) that contain identity

Sentence	True label
% of men lie to law-enforcement officers when question into domestic violence cases it is not without marks on the body or conclusive proof that the men are charged you sir you are one of the men that is the problem I can tell !	0
% of women lie to law-enforcement officers when question into domestic violence cases it is not without marks on the body or conclusive proof that the women are charged you mademoiselle you are one of the women that is the problem I can tell !	0

Table 11: Example of a sentence where the original target is a Male (top) and when the gender is swapped to Female (bottom).

	SenseScore		
Model	Gender	Race	Religion
AlBERT-base	$6.9e^{-05}$	0.032	0.006
+ downstream-perturbed-data	$\downarrow 4.2e^{-05}$	$\downarrow 0.002$	$\downarrow 0.001$
+ downstream-stratified-data	$\uparrow 0.042$	0.032	$\uparrow 0.009$
+ downstream-perturbed-stratified-data	$\uparrow 0.013$	$\downarrow 0.003$	$\downarrow 0.0007$
BERT-base	0.001	0.03	0.001
+ downstream-perturbed-data	$\downarrow 0.0007$	$\downarrow 0.003$	0.001
+ downstream-stratified-data	$\uparrow 0.025$	$\downarrow 0.022$	$\uparrow 0.004$
+ downstream-perturbed-stratified-data	$\uparrow 0.002$	$\downarrow 0.002$	$\downarrow 0.0008$
RoBERTa-base	0.001	0.024	0.003
+ downstream-perturbed-data	$\downarrow 0.0008$	$\downarrow 0.006$	$\downarrow 0.001$
+ downstream-stratified-data	$\uparrow 0.038$	$\uparrow 0.036$	0.003
+ downstream-perturbed-stratified-data	$\uparrow 0.003$	$\downarrow 0.002$	$\downarrow 0.0003$

Table 12: SenseScores of the difference models before and after the different debiasing methods. ($\uparrow$) denotes that the fairness metric score increased and the fairness worsened. ($\downarrow$) denotes that the fairness metric score decreased and the fairness improved.

($\hat{g}$), fairness is measured as:

$$SenseScore = |Mean_{x \in X}(f(\hat{x}) - f(x))| \tag{6}$$

$SenseScore$ is an indicator of how the model treats different groups of people, since the sentence is the same with only the identity group being different. The bigger the score, the less fair the model is, since it means that the model treats the different groups differently. On the contrary, the smaller the score, the more fair the model is, since it means that it does not discriminate between the different groups of people based on sensitive attributes. This analysis is possible because the balanced civil community fairness dataset (section 5.1Balanced fairness datasetsubsection.0.5.1) contains counterfactual/perturbed examples. Consequently, when we measured the $SenseScore$ of two identity groups, e.g., Male and Female, we are actually measuring the difference of the models' prediction probabilities between the same sentences with only the gender identity keywords being different. This analysis was conducted for the different downstream debiasing techniques that proved the most effective and improved the models' fairness without hurting the models' performance, which are: re-stratification (Downstream-stratified-data), perturbation (Downstream-perturbed-data), and re-stratification and perturbation (Downstream-perturbed-stratified-data).

We then examined the difference in $SenseScores$ of the examined debiasing techniques for the different sensitive attributes. For the gender attribute, we studied the sentences that are targeted at the Male group and that are perturbed to change the identity to the Female group, as shown in Table 11Example of a sentence where the original target is a Male (top) and when the gender is swapped to Female (bottom).table.0.11, as well as the sentences that are targeted at the Female group and are perturbed to change the identity to the Male group. Then, we measured the $SenseScore$ between the same sentences with the Male and the Female identities swapped. For the race attribute, we examined the sentences that are targeted at the Black group and that are perturbed to change the identity to the White group, as well as the sentences that are targeted at the White group and are perturbed to change the identity to the Black group. For the religion attribute, we examined the sentences that are targeted at the Christian group and are perturbed by changing the identity to the Muslim group, as well as the sentences that are targeted at the Muslim group and are perturbed to change the identity to the Christian group.

The prediction sensitivity scores ($SenseScore$) shown in Table 12SenseScores of the difference models before and after the different debiasing methods. ($\uparrow$) denotes that the fairness metric score increased and the fairness worsened. ($\downarrow$) denotes that the fairness metric score decreased and the fairness improved.table.0.12 indicate that removing overamplification bias is the most effective debiasing method. Fine-tuning the different models on a perturbed balanced dataset (+ downstream-perturbed-data) improved the fairness (lower $SenseScore$) for almost all the sensitive attributes, as evidenced by the results for AlBERT (gender, race, religion), BERT (gender, race), and RoBERTa (gender, race, religion). The next most effective debiasing method is removing both selection and overamplification bias, since fine-tuning the different models on a perturbed-re-stratified balanced dataset (+ downstream-perturbed-stratified-data) improved the fairness for all the models, but only for the race and the religion sensitive attributes, as evidenced by the results for AlBERT (race, religion), BERT (race, religion), and RoBERTa (race, religion). On the other hand, removing only the selection bias by fine-tuning the models on re-stratified data (+ downstream-stratified-data) is the least effective on the models' fairness, as it improved only the fairness of BERT (race) and deteriorated the fairness of AlBERT (gender, religion), BERT (gender, religion), and RoBERTa (gender, race).

8. How to improve fairness of Toxicity detection

Now we draw from our results and findings in this paper to answer our last research question and to provide a list of guidelines to ensure the fairness of the task of toxicity detection.

8.1 Guidelines

Based on the findings of this work, we recommend a list of steps to follow to ensure the fairness of the downstream task of toxicity detection, as examined in this work in the context of toxicity detection.

1. **Know the data:** The first recommendation is to know your data. This recommendation is the first step in any NLP task. But to ensure fairness, we also need to know about the bias in the training dataset. Especially since the results in Table 9The Pearson correlation coefficient (ρ) between different sources of bias and fairness metrics in all the models. **Bold** values refer to the highest (ρ) values and hence the strongest correlation.table.0.9 indicate that downstream sources of bias are the most influential on the models' fairness. We recommend measuring the selection bias and overamplification bias in the training dataset.

In addition to measuring selection and overamplification bias, it is important to generally, investigate annotators bias if available and to check the data collection process and other data properties as argues by Pushkarna, Zaldivar, and Kjartansson (2022). We also recommend learning about the historical and social context of the collected data to avoid undesirable societal impact (Benjamin 2019).

2. **Remove overamplification bias:** We recommend starting with removing the overamplification bias since it is the most impactful debiasing method on the models' fairness, as explained in section 7.3Which debiasing techniques to use to ensure the models' fairness on the task of toxicity detection?subsection.0.7.3. We recommend removing overamplification bias by fine-tuning the model on the perturbed dataset to ensure the balanced contextual representations of the different identity groups, as shown in section 5Fairness in the task of toxicity detectionsection.0.5. In addition to fine-tuning the model on balanced datasets, we recommend investigating different bias removal techniques to make sure that the right method is used with the right dataset.

3. **Know the model:** Similar to the data, it is important to know about the bias in the models being considered for the downstream task. Especially since the results suggest that there is a positive correlation between representation bias and the models' fairness. However, there are limitations related to the metrics used to measure intrinsic bias. We recommend using more than one metric. It is also important to check the intended use of the language model by checking the models' cards (Mitchell et al. 2019), if available.

4. **Balance the fairness data:** We recommend creating a perturbed version of the fairness dataset to make sure that the fairness dataset does not contain the selection or overamplification bias and that the measured fairness is more reliable, as explained in section 5Fairness in the task of toxicity detectionsection.0.5.

5. **Measure counterfactual fairness:** Since the different fairness metrics provide different results and sometimes are not in agreement, we recommend using counterfactual fairness metrics. Especially since after balancing the fairness dataset, we have perturbed data items. This allows us to reliably measure how the model discriminates or not between the different groups of people.

6. **Select the final model:** After collecting information about the bias in the fine-tuning dataset and the used language model and applying different debiasing methods to remove overamplification bias and measure fairness using counterfactual fairness metrics, we recommend to use the language model that achieves a good trade-off between performance and fairness.

To further assess the generalizability of these guidelines for the general task of supervised text classification, we applied these guidelines to ensure the fairness of the task of sentiment analysis, as discussed in Appendix 10Appendix H: Sentiment analysissection*.6, showing that these guidelines are generalizable. However, further analysis and experiments on larger datasets are needed, and we leave that for future work.

9. Limitations

It is important to point out that this work is limited to the examined models and datasets. Bias and fairness are studied from a Western perspective regarding language (English) and culture. In addition, some concerns have been raised regarding how the different bias and fairness metrics

articulate what the different metrics actually measure, as well as regarding the quality of the crowd-sourced datasets used to measure the bias (Blodgett et al. 2021). Besides, those metrics measure the existence of bias, not its absence, so a lower score does not necessarily mean that the model is unbiased (May et al. 2019). Nevertheless, when used for comparing different models, they can indicate whether a model is less biased than another. Additionally, we were not able to replicate the representation bias scores reported in Nangia et al. (2020) and (Nadeem, Bethke, and Reddy 2021), because the latest version of the Transformer's Python package that we used (version 4) led to different results compared to the one used by the authors (version 3). The same finding was made by Schick, Udupa, and Schütze (2021).

Despite the fairness metrics used in this work being the most commonly used ones in the literature, they are not without criticism (Hedden 2021). For example, Hedden (2021) argues that group fairness metrics are based on criteria that cannot be satisfied unless the models make perfect predictions or that the base rates (Bar-Hillel 1980) are equal across all the identity groups in the dataset.

Furthermore, we recognize that the provided recommendations to have a fairer toxicity detection task rely on creating perturbations for the training and the fairness dataset, and we acknowledge that this task may be challenging for some datasets, especially if the mention of the different identities is not explicit, like, for example, not using the word "Asian" to refer to an Asian person but using Asian names instead.

Moreover, in this work, we aim to achieve equity in the fairness of the task of toxicity detection between the different identity groups. However, equity does not necessarily mean equality, as explained in Broussard (2023).

10. Conclusion

This work examined the impact of three different sources of bias, Representation, Selection and Overamplification bias, on the fairness of the downstream task of toxicity detection. This study covers three sensitive attributes: gender, race, and religion, as well as three widely used language models (BERT-base, ALBERT-base, and RoBERTa-base). We used metrics from the literature to measure representation bias. And we proposed methods to measure selection and overamplification bias. We, then, measure the impact of the different sources of bias on the fairness of toxicity detection, and evaluated the impact of bias removal on improving the models' fairness and performance. Results showed that using a fairness dataset with the same contextual representation and ratio of positive examples for the different identity groups is a crucial step in measuring fairness. Unlike the findings of earlier research that suggested that there is no correlation between representation bias and the models' fairness, results show a consistent positive correlation between the representation bias scores measured by the CrowS-Pairs metric and fairness scores measured using different fairness metrics and when a balanced fairness dataset is used to measure fairness.

Additionally, our results consistently confirm that downstream sources of bias (selection and overamplification) are more impactful than upstream sources (representation bias), which is in line with other researchers' findings. Overamplification bias was shown to be the most impactful source of bias on the models' fairness, and removing it improved the fairness of the different models on the task of toxicity detection. Fine-tuning the models on the perturbed dataset led to the models providing similar prediction probabilities to the sentences where only the identity group is swapped, which suggests that the models do not discriminate between the different groups of people within the same identity group.

Finally, the findings of this work were used to devise a set of guidelines for ensuring the fairness of the downstream task of toxicity detection, in the hopes that they would assist researchers and data scientists in creating fairer toxicity detection models.

Female	she, her, hers, mum, mom, mother, daughter, sister, niece, aunt, grandmother, lady, woman, girl, ma'am, female, wife, ms, miss, mrs, ms., mrs.
Male	he, him, his, dad, father, son, brother, nephew, uncle, grandfather, gentleman, man, boy, 'ir, male, husband, mr, mr.
Neutral	they, them, theirs, parent, child, sibling, person, spouse

Table 1: Keywords used to filter out the gendered sentences in IMDB dataset.

Appendix A: Sentiment analysis

To train a sentiment analysis model that is fair, we first need a sentiment analysis dataset that contains information about sensitive attributes. Since, to the best of our knowledge, there is no sentiment analysis dataset that contains identity information, we filtered the IMDB sentiment dataset[1] to ensure that it contains gender information similar to the work in Webster et al. (2020). The keywords from Table 1Keywords used to filter out the gendered sentences in IMDB dataset.table.0.1 were used to filter the IMDB dataset to ensure that gendered information is present in the training dataset. A gender column was then added to the dataset, and sentences were labelled as "Female" or "Male" based on the gendered keywords. There are cases when none of the identity words were found or a mixture of the words was found in the same sentence. In such cases, these sentences were labelled as "Neutral". The IMDB dataset, after the keyword filtering, contained 50K data items. However, we then selected only the data items labelled as "Male" or "Female". We call the filtered dataset "IMDB-gendered" and it contains 9,790 data items. 72% of the data is labelled as "Male" and 27% is labelled as "Female". The ratio of the positive examples in the subset of sentences that contain the Male identity is 0.55 and the ratio of positive examples in the subset of sentences that contain the Female identity is 0.52. IMDB-gendered was then randomly split as 40% for training, 30% for validation, and 30% for the test, preserving the class ratio. Finally, three models were trained on the IMDB-gendered dataset, AlBERT-base, BERT-base, and RoBERTa-base, achieving AUC scores of 0.899, 0.912, and 0.914, respectively.

To measure gender fairness in the downstream task of sentiment analysis, we need a fairness dataset where the target of the sentiment in the sentence is identified. For this task, the SST-sentiment-fairness dataset (gero et al. 2023) was used. The dataset is a subset of the SST dataset that contains 462 data items with the target of the sentiment labelled by 3 annotators with an inter-annotation agreement of 0.65 and a Fleiss's Kappa score of 0.53. 42% of the dataset is targeted at women (ratio of positive examples = 0.61) and 58% is targeted at men (ratio of positive examples= 0.5). The three sentiment analysis models that we fine-tuned on the IMDB-gendered dataset were used to predict the labels of the SST-sentiment-fairness dataset and to measure their fairness, achieving AUC scores of 0.865 for AlBERT-base, 0.860 for BERT-base, and 0.878 for RoBERTA-base.

We then applied the fairness guidelines described in section 8.1Guidelinessubsection.0.8.1 to the sentiment analysis task as follows:

1. **Know the data:** When we applied this to the IMDB-gendered dataset, using the methods proposed in sections 6.3Selection biassubsection.0.6.3 and 6.5Overamplification biassubsection.0.6.5, we found that the selection bias score wass 0.03 and overamplification bias score was 0.309[2]

[1] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[2] Overamplification bias here is measured as the difference between the size of the male and female subsets divided by the size of the dataset.

2. **Remove overamplification bias:** We applied this technique to the IMDB-gendered dataset, and then, fine-tuned the models, AlBERT, BERT, and RoBERTa, on the perturbed-IMDB-gendered dataset. The AUC scores of the models on the perturbed-IMDB-gendered dataset are 0.869, 0.860, and 0.877 for AlBERT, BERT, and RoBERTa respectively.

3. **Know the model:** Since we already measured the representation bias in section 6.1Representation biassubsection.0.6.1, we do not need to repeat that for the sentiment analysis case study.

4. **Balance the fairness data:** We used the same method used in section 5Fairness in the task of toxicity detectionsection.0.5 to create the gendered perturbed SST-fairness dataset. 50% of the dataset is targeted at women (ratio of positive examples = 0.54) and 50% is targeted at men (ratio of positive examples = 0.54). The data after perturbation contained 924 data items, with selection bias and overamplification bias both being 0.

5. **Measure counterfactual fairness:**For the sentiment analysis task, we measured the fairness of the models after being fine-tuned on the perturbed IMDB-gendered dataset. We measured the counterfactual fairness of the models AlBERT, BERT, and RoBERTa on the perturbed SST-sentiment-fairness dataset, as explained in section 7.3Which debiasing techniques to use to ensure the models' fairness on the task of toxicity detection?subsection.0.7.3. We found that the counterfactual fairness, as measured using the $Sensescore$ of the different models after removing Overamplification bias, is 0.005 for AlBERT, 0.0005 for BERT-base, and 0.009 for RoBERTa-base.

6. **Select the final model:** The results indicate that after removing overamplification bias, the model that discriminates the least between the male and the female groups in our sentiment analysis task is BERT (couterfactual fairness of 0.0005). Consequently, we chose BERT as it is the fairest model to use. This decision comes with a trade-off with regards to the performance score. On the perturbed-gendered-IMDB dataset, RoBERTa (+ downstream-perturbed-data) achieves an AUC score of 0.877, which slightly outperforms BERT (+ downstream-perturbed-data) that achieved an AUC score of 0.860.

Acknowledgments

This research has been partially supported by project "AGENCY", funded by the Engineering and Physical Sciences Research Council [EP/W032481/1], United Kingdom. Part of this work was done when the first author did an internship at IBM research in New York, the US.

References

Abdi, Hervé and Lynne J Williams. 2010. Principal component analysis.(2010). *Computational Statistics, John Wiley and Sons*, pages 433–459.

AbdulNabi, Isra'a and Qussai Yaseen. 2021. Spam email detection using deep learning techniques. *Procedia Computer Science*, 184:853–858. The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.

Badilla, Pablo, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: the word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 430–436, ijcai.org.

Baldini, Ioana, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Association for Computational Linguistics, Dublin, Ireland.

Bar-Hillel, Maya. 1980. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233.

Benjamin, Ruha. 2019. *Race after technology: Abolitionist tools for the new jim code*. Polity.

Blodgett, Su Lin, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna M. Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1004–1015, Association for Computational Linguistics.

Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Broussard, Meredith. 2023. *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cao, Yang, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Association for Computational Linguistics, Dublin, Ireland.

Caton, Simon and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

De-Arteaga, Maria, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128, ACM.

Dev, Sunipa and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887, PMLR.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186, Association for Computational Linguistics.

Elsafoury, Fatma. 2023. Systematic offensive stereotyping (sos) bias in language models. *ArXiv*, abs/2308.10684.

Elsafoury, Fatma, Stamos Katsigiannis, Steven R. Wilson, and Naeem Ramzan. 2021. Does BERT pay attention to cyberbullying? In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1900–1904, Association for Computing Machinery, New York, NY, USA.

Elsafoury, Fatma, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, International Committee on Computational Linguistics, Gyeongju, Republic of Korea.

Elsafoury, Fatma, Steven R. Wilson, and Naeem Ramzan. 2022. A comparative study on word embeddings and social NLP tasks. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Association for Computational Linguistics, Seattle, Washington.

Fryer, Zee, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, and Kellie Webster. 2022a. Flexible text generation for counterfactual fairness probing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 209–229, Association for Computational Linguistics, Seattle, Washington (Hybrid).

Fryer, Zee, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, and Kellie Webster. 2022b. Flexible text generation for counterfactual fairness probing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 209–229, Association for Computational Linguistics, Seattle, Washington (Hybrid).

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Garg, Sahaj, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 219–226, ACM.

Garrido-Muñoz, Ismael, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7).

gero, Katy, Nathan Butters, Anna Bethke, and Fatma Elsafoury. 2023. A dataset to measure fairness in the sentiment analysis task. https://github.com/efatmae/SST_sentiment_fairness_data.

Goldfarb-Tarrant, Seraphina, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Association for Computational Linguistics, Online.

Gonen, Hila and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Guo, Wei and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, Association for Computing Machinery, New York, NY, USA.

Hartvigsen, Thomas, 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, Association for Computational Linguistics, Dublin, Ireland.

Hedden, Brian. 2021. On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs*, 49(2):209–231.

Hovy, Dirk and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Hutchinson, Ben and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 49–58, Association for Computing Machinery, New York, NY, USA.

Jourdan, Fanny, Laurent Risser, Jean-michel Loubes, and Nicholas Asher. 2023. Are fairness metric scores enough to assess discrimination biases in machine learning? In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 163–174, Association for Computational Linguistics, Toronto, Canada.

Kaneko, Masahiro, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, International Committee on Computational Linguistics, Gyeongju, Republic of Korea.

Krishna, Satyapriya, Rahul Gupta, Apurv Verma, J. Dhamala, Yada Pruksachatkun, and Kai-Wei Chang. 2022. Measuring fairness of text classifiers via prediction sensitivity. *ArXiv*, abs/2203.08670.

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

Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net.

li, Weijun and Zhenyu Liu. 2011. A method of svm with normalization in intrusion detection. *Procedia Environmental Sciences*, 11:256–262. 2011 2nd International Conference on Challenges in Environmental Science and Computer Engineering (CESCE 2011).

Liang, Paul Pu, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Association for Computational Linguistics, Online.

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

Lopez, Paola. 2021. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, 10(4):1–29.

Ma, Edward. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Mathew, Binny, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

May, Chandler, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 622–628, Association for Computational Linguistics.

Meade, Nicholas, 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, Association for Computational Linguistics, Dublin, Ireland.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, Association for Computing Machinery, New York, NY, USA.

Mollas, Ioannis, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Nadeem, Moin, 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, Association for Computational Linguistics, Online.

Nangia, Nikita, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Association for Computational Linguistics, Online.

Ng, Jun-Ping and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Nozza, Debora, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Association for Computational Linguistics, Dublin, Ireland.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*, Association for Computational Linguistics.

Ousidhoum, Nedjma, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained 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 4262–4274, Association for Computational Linguistics, Online.

Papakipos, Zoe and Joanna Bitton. 2022. Augly: Data augmentations for robustness.

Prabhakaran, Vinodkumar, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model 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 5740–5745, Association for Computational Linguistics, Hong Kong, China.

Pushkarna, Mahima, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, Association for Computing Machinery, New York, NY, USA.

Qian, Rebecca, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Association for Computational Linguistics, Florence, Italy.

Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Association for Computational Linguistics, Online.

Schick, Timo, 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.

Seshadri, Preethi, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*.

Shah, Deven Santosh, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Association for Computational Linguistics, Online.

Steed, Ryan, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Association for Computational Linguistics, Dublin, Ireland.

Sweeney, Chris and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Association for Computational Linguistics, Florence, Italy.

Webster, Kellie, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. Technical report, Google Research.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Association for Computational Linguistics, Copenhagen, Denmark.

Zhu, Yutao, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. 2023. Large language models for information retrieval: A survey. *ArXiv*, abs/2308.07107.

Zmigrod, Ran, 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, Association for Computational Linguistics, Florence, Italy.