# Causal ATE Mitigates Unintended Bias in Controlled Text Generation

**Rahul Madhavan**
IISc, Bangalore
mrahul@iisc.ac.in

**Kahini Wadhawan**
IBM Research, Delhi
kahini.wadhawan1@ibm.com

## Abstract

We study attribute control in language models through the method of Causal Average Treatment Effect (Causal ATE). Existing methods for the attribute control task in Language Models (LMs) check for the co-occurrence of words in a sentence with the attribute of interest, and control for them. However, spurious correlation of the words with the attribute in the training dataset, can cause models to hallucinate the presence of the attribute when presented with the spurious correlate during inference. We show that the simple perturbation-based method of Causal ATE removes this unintended effect. Specifically, we ground it in the problem of toxicity mitigation, where a significant challenge lies in the inadvertent bias that often emerges towards protected groups post detoxification. We show that this unintended bias can be solved by the use of the Causal ATE metric. We provide experimental validations for our claims and release our code (anonymously) here: github.com/causalate-mitigates-bias/causal-ate-mitigates-bias.

## 1 Introduction

Controllable text generation methods are often used to guide the text generated by language models (LMs) towards certain desirable attributes (Hu and Li, 2021; Dathathri et al., 2019; Liu et al., 2021). The goal herein is to generate sentences whose attributes can be controlled (Prabhumoye et al., 2020). Language models, which are pre-trained only for next word prediction, cannot directly control for attributes in their outputs. On the other hand, one may wish to alter words in the autoregressively produced sentences, either accentuating or mitigating the desired attributes. Attributes such as sentiment, writing style, language precision, tone, and toxicity are key concerns for control in language models, with particular emphasis on toxicity mitigation due to its relevance in sensitive contexts (Perez et al., 2020).

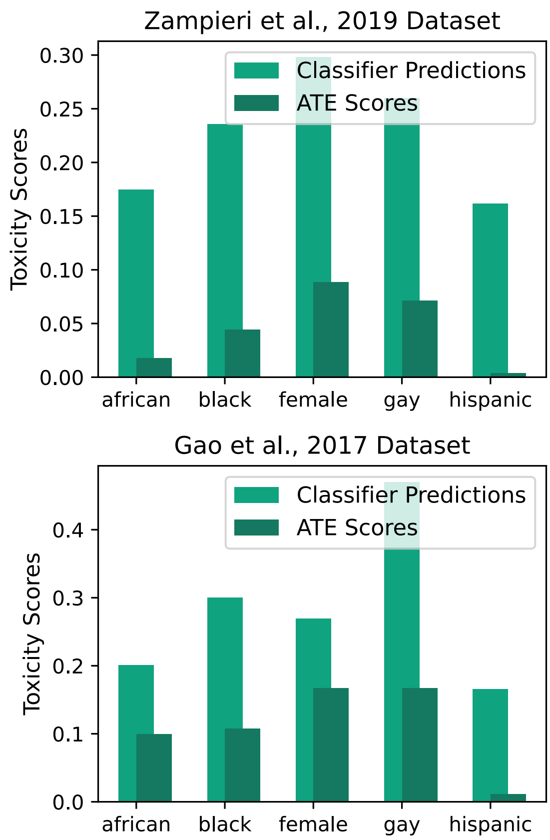


Figure 1: We plot the ATE score vs a regression based classifier for toxicity across two datasets. ATE Scores show a lower toxicity for protected groups.

Regularizers in the reward models are often employed during training to alter the output sentences towards certain desirable attributes (Hu et al., 2017). Such regularization penalities (or rewards) often rely on models trained on real-world datasets. Such datasets contain spurious correlates – words that correlate with certain attributes without necessarily causing them (Nam et al., 2020; Udomcharoenchaikit et al., 2022).

In the context of toxicity mitigation, prior works show that detoxification methods inadvertently impact language model outputs concerning marginal-



*Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 130–142
November 15-16, 2024 

ized groups (Welbl et al., 2021). Words such as 'gay' or 'female' are identified as being toxic, as they co-occur with toxic text, and hence the LM stops speaking about them (Xu et al., 2021).

This is called the *unintended bias problem*. In this paper we provide experimental and theoretical justifications for the use of causal ATE to mitigate the unintended bias problem in text classification. We prove theoretically that for spurious correlates, the causal ATE score is upper-bounded. We also show through extensive experiments on two popular toxicity classification datasets (Zampieri et al., 2019a; Gao and Huang, 2017) that our method shows experimental promise (See Figure 1).

We provide a full list of related works in Related Works section 6.

### 1.1 Our Contributions:

**1.** We show theoretically that the Causal ATE score of spurious correlates is less than $0.25$ under mild assumptions in Sections 2 and 3.

**2.** We provide a theoretical basis for the study of the perturbation based Causal ATE method. We show that it can be used alongside any classifier towards improving it for false positive rates.

**3.** We provide experimental validation for our claims by showing that causal ATE scores indeed decrease the toxicity for spurious correlates to toxic sentences in Section 4.

## 2 Notations and Methodology

Consider a sentence $s$, made up of tokens (words) from some universe of words $W$. Let the list of all sentences $s$ in our dataset be denoted $\mathcal{S}$. Let each sentence $s \in \mathcal{S}$ be labelled with the presence or absence of an attribute $A$. So the dataset, which we can call $\mathcal{D}$, consists of tuples $(s, A(s))$ for all $s \in \mathcal{S}$. Let the cardinality of the labelled dataset be $|\mathcal{D}| = |\mathcal{S}| = n$.

From such a dataset, it is possible to construct an attribute model that gives us an estimate of the probability of attribute $A$, given a sentence $s$. i.e. It is possible to construct a model $\widehat{A}(\cdot)$ such that $\widehat{A}(s) = \widehat{\mathbb{P}}\{A \mid s\}$ for any given sentence $s$. Now such a model may rely on the words in $s$. Let $s = \{w_1, \dots, w_n\}$. We now define an attribute model $\widehat{a}(\cdot)$ given a word as follows:

**Definition 1** (Attribute model $\widehat{a}(w_i)$ for any word $w_i \in W$)**.**

$$\widehat{a}(w_i) := \frac{|\{\text{sentences } s \in \mathcal{D} \text{ containing } w_i \text{ s.t. } A(s) = 1\}|}{|\{\text{sentences } s \in \mathcal{D} \text{ containing } w_i\}|} \tag{1}$$

$$= \frac{n(A(s) = 1 \mid w_i \in s)}{n(s \mid w_i \in s)} \tag{2}$$

where $n(\cdot)$ denotes the cardinality of the set satisfying the properties.

Note that such a model is purely correlation based, and can be seen as the proportion of sentences containing an attribute amongst those containing a particular word. i.e. it is an estimate of the co-occurrence of attribute with the word. Based on attribute model $\widehat{a}(\cdot)$ we can define an attribute model $\widehat{A}(\cdot)$ for any sentence $s = \{w_1, \dots, w_k\}$ as follows:

**Definition 2** (Attribute model $\widehat{A}(s)$ for a sentence $s \in W^k$)**.**

$$\widehat{A}(s = \{w_1, \dots, w_k\}) := \max_{w_i \in s} \widehat{a}(w_i) \tag{3}$$

$$= \max\{\widehat{a}(w_1), \dots, \widehat{a}(w_k)\} \tag{4}$$

Note that such a model is conservative and labels a sentence as having an attribute when any word in the sentence has the attribute. For the purpose of attributes such as toxicity, such an attribute model is quite suitable.

### 2.1 Computation of ATE Score of a word with respect to an attribute

Given a model representing the estimate of the attribute $A$ in a sentence $s$, denoted as $\widehat{\mathbb{P}}\{A(s) = 1\}$, we can now define the ATE score. Note that the Causal ATE score does not depend on the particular model for the estimate $\widehat{\mathbb{P}}\{A(s) = 1\}$ – i.e. we can use any estimator model.

If we denote $f_A(s)$ as the estimate of $\mathbb{P}\{A(s) = 1\}$ obtained from *some* model. We can then define Causal ATE with respect to this estimate. If a sentence $s$ is made up of words $\{w_1, \dots, w_i, \dots, w_k\}$. For brevity, given a word $w_i$, from a sentence $s$, we may refer to the rest of the words in the sentence as context $c_i$. Consider a *counter-factual* sentence $s'$ where (only) the $i$th word is changed: $\{w_1, \dots, w_i', \dots, w_k\}$. Such a word $w_i'$ may be the most probable token to replace $w_i$, given the rest of the sentence.

We now define a certain value that may be called the Treatment Effect (TE), which computes the effect of replacement of $w_i$ with $w_i'$ in sentence $s$, on the attribute probability.

**Definition 3** (Treatment Effect (TE) of a word in a sentence given replacement word)**.** Let word $w_i$ be replaced by word $w'_i$ in a sentence $s$. Then:

$$\begin{aligned}\mathsf{TE}(s, w_i, w'_i) &= f_A(s) - f_A(s') \\ &= f_A(\{w_1, \ldots, w_i, \ldots, w_k\}) \\ &\quad - f_A(\{w_1, \ldots, w'_i, \ldots, w_k\})\end{aligned} \tag{5}$$

The expectation now can be taken over the replacement words, given the context, and over all contexts where the words appear.

**Definition 4** ($\mathsf{ATE}$ of word $w_i$ given dataset $\mathcal{D}$ and an attribute classifier $f(\cdot)$)**.**

$$\mathsf{ATE}(w_i) = \mathop{\mathbb{E}}_{s \in \mathcal{D} | w_i \in s} \left[ f(s) - \mathop{\mathbb{E}}_{w'_i \in W} [f(s')] \right] \tag{6}$$

where $s'$ is the sentence $s$ where word $w_i$ is replaced by $w'_i$

This $\mathsf{ATE}$ score precisely indicates the intervention effect of $w_i$ on the attribute probability of a sentence. Notice that this score roughly corresponds to the *expected difference in attribute on replacement* of word.

Now say we compute the $\mathsf{ATE}$ scores for every token $w$ in our universe $W$ in the manner given by Equation 6. We can store all these scores in a large lookup-table. Now, we are in a position to compute an attribute score given a sentence.

### 2.2 Computation of Attribute Score for a sentence

The causal ATE approach suggests that we can build towards the ATE of a sentence given the ATE scores of each of the words in the sentence recursively. We illustrate this approach in Figure 2. First, note that each word $w_t$ is stochastically generated based on words $w_1, \ldots, w_{t-1}$ in an auto-regressive manner. If we denote $\{w_1, \ldots, w_{t-1}\}$ as $s_{t-1}$, then we can say the distribution for $w_t$, is generated from $s_{t-1}$ and the structure of the language. To sample from the probabilistic distribution, we may use an exogenous variable such as $U_t$.

The attribute $A(s_{t-1})$ of a sentence up to $t-1$ tokens, depends only on $\{w_1, \ldots, w_{t-1}\} \equiv s_{t-1}$. We now describe a model for computing attribute $A(s_t)$ from $A(s_{t-1})$ and $\mathsf{ATE}(w_t)$. The larger English causal graph moderates influence of $w_t$ on $A(s_t)$ through the $\mathsf{ATE}$ score of the words. We consider $A(s_t) = \max(A(s_{t-1}), \mathsf{ATE}(w_t))$. This is equivalent to

$$A_\infty(s = \{w_1, \ldots, w_n\}) = \max_{i \in [n]} \mathsf{ATE}(w_i) \tag{7}$$

More generally, we propose an attribute score $A(s)$ for this sentence given by $A(s) = \|\{\mathsf{ATE}(w_1), \ldots, \mathsf{ATE}(w_n)\}\|_p$ where $\|\cdot\|_p$ indicates the $L_p$-norm of a vector. We can call these attribute scores $A(s)$ as the $\mathsf{ATE}$ scores of a sentence.

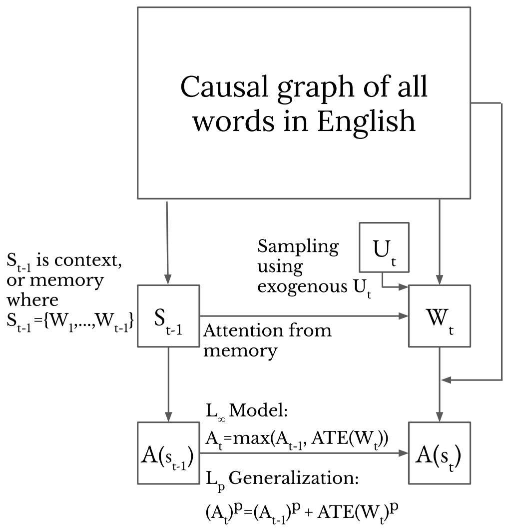


Figure 2: An Illustration of the Causal Graph used to compute the attribute score of a sentence recursively.

## 3 Theory and Background

Now that we have laid the groundwork, we can make proceed to make the central claims of this work.

**Lemma 1.** Consider sentence $s = \{w_1, \ldots, w_k\}$. We will make two simple claims:

1. If $\nexists w_i \in s$ such that $\mathsf{ATE}(w_i) \geq c$, then, $A(s) < c$.
2. If $\exists w_i \in s$ such that $\mathsf{ATE}(w_i) \geq c$, then, $A(s) \geq c$.

This lemma is straightforward to prove from Definition 7.

We will now make a claim regarding the $\mathsf{ATE}$ score of the given words themselves. Recall that $c_i$ is the context for the word $w_i$ from a sentence $s$. Given $c_i$, $w_i$ is replaced by $w'_i$ by a perturbation model (through Masked Language Modelling). Towards our proof, we will make two assumptions:

Table 1: Description of Classifiers Used in Experiments

| Sl. No. | Model | Description |
|---|---|---|
| 1 | Logistic Regression (LR) | A linear classifier that predicts toxicity using logistic regression. |
| 2 | SVM | Support Vector Machine with a linear kernel for text classification. |
| 3 | Gradient Boosting (GB) | An ensemble model that combines weak learners for enhanced toxicity prediction. |
| 4 | Naive Bayes (NB) | Multinomial Naive Bayes, a probabilistic model for text classification. |
| 5 | NN1Layer5 | Neural network with 1 hidden layer of 5 neurons. |
| 6 | NN2Layer105 | Neural network with 2 hidden layers (10 neurons and 5 neurons, respectively). |
| 7 | NN3Layer20105 | Neural network with 3 hidden layers (20, 10, and 5 neurons, respectively). |

*Assumption* 1. We make a mild assumption on this replacement process: $\widehat{a}(w_i') < \widehat{A}(c_i)$. Grounding this in the attribute of toxicity, we can say that the replacement word is less toxic than the context. This is probable if the replacement model has been trained on a large enough corpus. See (Madhavan et al., 2023) for empirical results showing this claim to be true in practice.

*Assumption* 2. We make an assumption on the dataset. A *spurious correlate* has a word with a higher attribute score in the rest of the sentence for sentences labelled as having the attribute. For example, in the case of toxicity, a spurious correlate like Muslim, has a more toxic word in the rest of the sentence, when the sentence is labelled as toxic.

Given these assumptions, we have the following theorem:

**Theorem 1.** Given Assumptions 1 and 2 for a spurious correlate $w_i$, $\texttt{ATE}(w_i) \leq 0.25$.

*Proof.* If we consider three numbers $\{\widehat{A}(c_i), \widehat{a}(w_i), \widehat{a}(w_i')\}$, there are six possible orderings of this set. We can subsume these orderings into two cases:

1. $\widehat{A}(c_i) < \widehat{a}(w_i')$.
2. $\widehat{A}(c_i) \geq \widehat{a}(w_i')$.

Within these cases, we study the variation of $ATE(w_i)$ with $\widehat{a}(w_i)$. We plot these in the Figure 3. Using a case-by-case analysis over these possibilities, we prove the statement.

The full proof of the Theorem is provided in Appendix A.

□

Based on Theorem A and Lemma 1, $A(s) \leq 0.25$ if each $w_i \in s$ is a spurious correlate, i.e. non-causal, for attribute $A$.

In the following section we provide experimental justification for our work through experimental results.

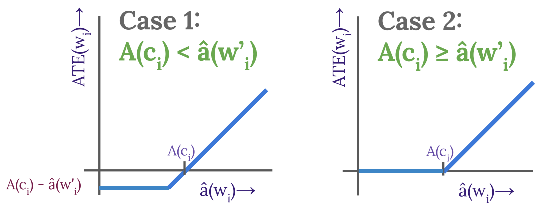


Figure 3: Graph of ATE score of a given word $w_i$ with $\widehat{a}(w_i)$ given two cases

## 4 Experiments

In this section, we present experimental evidence demonstrating the efficacy of the Causal Average Treatment Effect (Causal ATE) method for mitigating unintended bias in text classification tasks. Our experiments focus on toxicity detection, utilizing two widely recognized datasets. The results provide both theoretical and practical support for the utility of Causal ATE in addressing bias associated with protected groups.

### 4.1 Datasets and Preprocessing

We conducted experiments using two well-known datasets: the SemEval dataset (Zampieri et al., 2019a) and the dataset from Gao et al. (Gao and Huang, 2017). The SemEval dataset consists of tweets annotated for offensive language, while the Gao et al. dataset comprises user comments from Yahoo! News articles labeled for hate speech and harassment. These datasets were chosen for their diverse and challenging nature, providing an ideal testbed for evaluating bias mitigation in toxicity classification tasks.

The data was preprocessed to clean the text by removing special characters, URLs, and stop words. We used the `CountVectorizer` from the `scikit-learn` library to convert the textual data into a Bag-of-Words representation, ensuring a structured and uniform input for the classifiers.

Table 2: ATE Scores vs Classifier Predictions for different models by Protected Category for the Gao et al. Dataset

| Group → | **African** | | | **Black** | | | **Female** | | | **Gay** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model ↓ | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff |
| LR | 0.201 | 0.099 | 0.102 | 0.300 | 0.108 | 0.192 | 0.270 | 0.167 | 0.103 | 0.470 | 0.167 | 0.303 |
| SVM | 0.282 | 0.062 | 0.220 | 0.282 | 0.052 | 0.230 | 0.301 | 0.082 | 0.219 | 0.371 | 0.154 | 0.217 |
| GB | 0.225 | 0.052 | 0.173 | 0.335 | 0.071 | 0.264 | 0.225 | 0.000 | 0.225 | 0.653 | 0.204 | 0.449 |
| NB | 0.460 | 0.002 | 0.458 | 0.510 | 0.047 | 0.463 | 0.444 | 0.004 | 0.440 | 0.657 | 0.107 | 0.550 |
| NN1Layer5 | 0.000 | 0.003 | -0.003 | 0.000 | 0.059 | -0.059 | 0.000 | 0.024 | -0.024 | 1.000 | 0.197 | 0.803 |
| NN2Layer105 | 0.000 | 0.000 | 0.000 | 0.000 | 0.096 | -0.096 | 0.002 | 0.000 | 0.002 | 1.000 | 0.217 | 0.783 |
| NN3Layer20105 | 0.000 | 0.160 | -0.160 | 0.000 | 0.097 | -0.097 | 0.000 | 0.000 | 0.000 | 0.993 | 0.165 | 0.828 |

Table 3: ATE Scores vs Classifier Predictions for different models by Protected Category for the Zampieri et al. Dataset

| Group → | African | | | Black | | | Female | | | Gay | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model ↓ | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff |
| LR | 0.174 | 0.020 | 0.154 | 0.236 | 0.049 | 0.187 | 0.297 | 0.075 | 0.223 | 0.260 | 0.098 | 0.162 |
| SVM | 0.248 | 0.030 | 0.218 | 0.267 | 0.036 | 0.232 | 0.337 | 0.068 | 0.269 | 0.265 | 0.033 | 0.232 |
| GB | 0.269 | 0.020 | 0.249 | 0.269 | 0.013 | 0.256 | 0.269 | 0.008 | 0.261 | 0.269 | 0.003 | 0.266 |
| NB | 0.349 | 0.009 | 0.341 | 0.453 | 0.055 | 0.398 | 0.343 | 0.183 | 0.160 | 0.539 | 0.070 | 0.469 |
| NN1Layer5 | 0.000 | 0.000 | -0.000 | 0.000 | 0.052 | -0.052 | 0.000 | 0.000 | -0.000 | 0.000 | 0.114 | -0.114 |
| NN2Layer105 | 0.000 | 0.000 | 0.000 | 0.000 | 0.090 | -0.090 | 0.000 | 0.170 | -0.170 | 0.000 | 0.104 | -0.104 |
| NN3Layer20105 | 0.000 | 0.200 | -0.200 | 0.000 | 0.126 | -0.126 | 0.000 | 0.075 | -0.075 | 0.000 | 0.046 | -0.046 |

This vectorized representation was then used as input for the various models described in the next section.

### 4.2 Classifiers

We trained several classifiers to predict toxicity in sentences. These classifiers span traditional machine learning models and modern neural networks, allowing us to evaluate bias mitigation across a range of approaches. Table 1 provides a summary of the classifiers used in our experiments.

These models were implemented using the `scikit-learn` library. For the neural networks, we used the `MLPClassifier` with the `lbfgs` solver and a maximum of 10,000 iterations to ensure convergence during training.

### 4.3 Computation of ATE Scores

For each classifier, we computed the Causal ATE scores for a set of bias-inducing words related to protected groups, including "female", "black", "gay", "hispanic", and "african". These scores were calculated using a perturbation-based approach, where we replaced specific words in a sentence with alternatives generated by a masked language model (`roberta-base`). The ATE score measures the expected change in toxicity prediction when a particular word is replaced, providing insight into the causal effect of each word on the classifier's output.

This process enabled us to quantify the impact of potentially bias-inducing terms, allowing for a more nuanced understanding of how certain words contribute to biased predictions. By analyzing these ATE scores, we could identify instances where the classifier was overly reliant on spurious correlations, thus flagging cases of unintended bias.

### 4.4 Implementation and Runtime Considerations

The implementation of the experiments was carried out using `scikit-learn` for classifier training and the `transformers` library for masked token replacements using `roberta-base`. To ensure reproducibility, all experiments were conducted with a fixed random seed. The runtime for the entire experiment, when the preprocessed data files were available, was approximately 40 minutes on a single CPU thread. The most computationally expensive tasks included training the classifiers and generating the masked replacements for the ATE computations.

The code for our experiments, including data preprocessing, model training, and ATE computations, is available in our anonymous GitHub repository: github.com/causalate-mitigates-bias/causal-ate-mitigates-bias.

### 4.5 Discussion

From the results, we observe the following:
**1. Reduction in Predicted Toxicity**: The ATE scores are consistently lower than the original predicted probabilities for most classifiers and protected categories. This indicates that the Causal ATE method effectively reduces the unintended bias towards these groups.
**2. Classifier Performance Variance**: Naive Bayes (NB) shows the highest predicted probabilities and substantial differences (**Diff**) across all categories, suggesting a strong sensitivity to spurious correlations. In contrast, Neural Network models often exhibit lower predicted probabilities but sometimes result in negative **Diff** values, indicating overcorrection or model underfitting.
**3. Impact on Protected Categories**: Categories like *"Gay"* and *"Black"* show significant reductions in toxicity scores after applying the Causal ATE method. This aligns with our objective of mitigating bias towards marginalized groups.
**4. Consistency Across Datasets**: Similar trends are observed in both datasets, reinforcing the robustness of the Causal ATE approach in different contexts.

### 4.6 Conclusion of Experiments

The experimental results validate our theoretical claims that the Causal ATE method is an effective approach to mitigate unintended bias in toxicity classification tasks. By focusing on the causal impact of words rather than their spurious correlations, the method significantly reduces bias toward protected groups. Our experiments demonstrate that this approach is robust across different classifiers and datasets, offering a promising solution to bias mitigation in language models.

## 5 Discussion

### 5.1 Causal ATE is Generalizable

While we our experimental results have pertained to the use of Causal ATE as a metric for mitigating bias in toxicity classification, our theoretical results extend to any language attributes.

Figure 4 showcases different style attributes to which such an analysis can be applied. We hope that such causal approaches can be utilized for general use cases such as style control using LLMs.

While the main sections in the paper consider the attribute class of toxicity, we illustrate here that this method can equally be used for various attribute

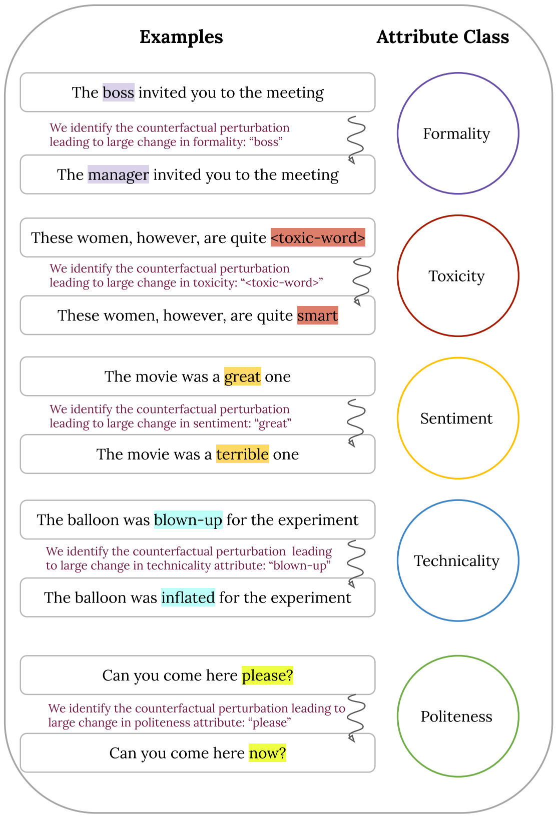


Figure 4: Illustration of word perturbation for identifying important words with respect to an attribute.

classes thereby easily scalable and generalizable. For instance, in the case of a style like formality, changing 'boss' to 'manager' has changes the sentence attribute to being more formal. Similarly, a change from the word 'terrific' or 'great' to 'terrible' in the context of a movie review, changes the entire meaning of a sentence, and effectively conveys a more negative sentiment.

Similarly, simple word changes can lead to the language being more technical or polite. Figure 4 illustrates that causal ATE can be used across various attributes for bias mitigation. The underlying idea is that we can perturb particular words in their context to check the change that they cause on the desired attribute.

### 5.2 Importance of using a Causal Graph

Given estimates of the probability $\mathbb{P}\{a_i \mid s\}$ for attributes in text generated by a Language Model (LM), the potential for fine-tuning the LM towards specific attributes becomes apparent. However, numerous challenges persist.
Firstly, attribute classifiers are prone to spurious correlations. For instance, if a protected token like

'Muslim' frequently appears in toxic sentences, the attribute classifier detecting toxicity might penalize the generation of the word 'Muslim'. This brings out in light that there is a trade-off between detoxification of LM and LM quality for text generation clearly detailed out in (Welbl et al., 2021). LM avoids to generate sentences containing protected tokens leading to higher perplexity for texts with these protected attrbiutes. Additionally, these classifier models providing $\mathbb{P}\, a_i \mid s$ estimates themselves may be LMs, resulting in slow training and requiring substantial computational resources.
Utilizing a causal graph directly addresses these challenges. It offer computational efficiency during training and are immune to spurious correlations, detecting interventional attribute distributions rather than conditional distributions through counterfactual interventions. Moreover, we get both flexibility and transparency regarding their exact form, features unavailable with LM classifiers.

## 6 Related Works

In this section we will look at five related lines of work: (a) Controlled generation (b) Unintended Bias problem (c) Toxicity Mitigation (d) Toxicity Detection (e) Causal Methods for Text

**Controlled Generation** can be broadly categorized into fine-tuning methods (Krause et al., 2020), data-based (Keskar et al., 2019; Gururangan et al., 2020), decoding-time approaches using attribute classifiers (Dathathri et al., 2019; Krause et al., 2020) and causality based approaches (Madhavan et al., 2023). Majority of these techniques were tested on toxicity mitigation and sentiment control. The dependence of attribute regularizers on probabilistic classifiers make them prone to such spurious correlations (Kaddour et al., 2022; Feder et al., 2022).
In the **Unintended Bias problem** LMs which are detoxified inherit a tendency to be biased against protected groups. LM quality is compromised due to a detoxification side-effect (Welbl et al., 2021; Xu et al., 2021). Some works address LM control through improving datasets (Sap et al., 2019b). Unfortunately, this makes annotation and data curation more expensive. As an alternative, there is growing interest in training accurate models in presence of biased data (Oren et al., 2019). Our work fits into this framework.
In the context of **Toxicity Mitigation**, (Welbl et al., 2021) highlight that detoxification methods have unintended effects on marginalized groups. They showcased that detoxification makes LMs more brittle to distribution shift, affecting its robustness in certain parts of language that contain mentions of minority groups. Concretely, words such as "female" are identified as being toxic, as they co-occur with toxic text, and hence the LM stops speaking about them (Xu et al., 2021). This is called the unintended bias problem. This unintended bias problem can manifest as differences in performance of the LM for different demographic groups.
**Toxicity Detection** Toxicity is a well studied problem in context of responsible and safe AI effort. Hence, we foucs our experiments on toxicty mitigation in this study. Several works have also studied the angle from toxic text detection. Numerous studies have explored toxic text detection, including HATEBERT (Caselli et al., 2020), HATECHECK (Röttger et al., 2020), and PERSPECTIVE API (Lees et al., 2022). We employ the HATEBERT model for assessing local hatefulness and utilize PERSPECTIVE API for third-party evaluation, where we report the corresponding metrics.

**Causal Methods for Text** Spurious correlations between protected groups and toxic text can be identified is by understanding the causal structure. (Feder et al., 2022) emphasizes on the connect between causality and NLP. Towards mitigation of the bias problem (Madhavan et al., 2023) proposed the use of Causal ATE as a regularization technique and showed experimentally that it does indeed perform as intended.

In this paper, we probe the Causal ATE metric theoretically, and prove that the Causal ATE metric is less susceptible to false positives. An attribute control method based on this metric would mitigate unintended bias. We provide a theoretical basis from which to understand the Causal ATE metric and showcase that this causal technique provides robustness across contexts for attribute control in language models.

## 7 Conclusion

In conclusion, our work provides a theoretical justification for using the causality-based concepts of counterfactuals, and ATE scores for controlled text generation. We provide experimental results that validate these claims. We show that the simple perturbation-based method of Causal ATE removes the unintended bias effect through reduction of false positives, additionally making systems more robust to biased data.

## 8 Limitations

The limitations of our proposed framework are described in detail in this section.

**1. Owing to Pre-trained models:** Third-party hatespeech detectors such as HATEBERT tend to overestimate the prevalence of toxicity in texts having mentions of minority or protected groups due to sampling bias, or just spurious correlations (Paz et al., 2020; Waseem, 2016; Dhamala et al., 2021). ATE computation though following causal mechanisms rely on these detectors for initial attribute probability scores. Additionally, these models suffer from low annotator agreement during dataset annotation because of absence of concrete defining hatespeech taxonomy (Sap et al., 2019a). Causal nature of our approach tends to mitigates bias but not completely eliminated the problem.

**2. Owing to language and training corpus:** We showcase empirically the utility of our theoretical claims in this study and conducted monolingual experiments on English language which could be further extended to other languages. Additionally, training corpora used for training HATEBERT and MLM model are known to contain curated data from internet, where reliability and factual accuracy is a known issue (Gehman et al., 2020). Hence, we are limited by the distributions of our training corpora in terms of what the model can learn and infer.

**3. Owing to distribution shift between datasets:** There are limitations that get introduced due to change in vocabulary from training to test sets. Sometimes, words which occur in test set are not in ATE training set, we ignore such words but could impact downstream perfomance of LLM if word was important. In case of such a distribution shift between the datasets, our model may not work as expected.

## 9 Ethics Statement

Our paper addresses the crucial issue of bias and toxicity in language models by using causal methods that involve several ethical concerns, that we address herein:

**1. Monolingual limitation :** This work addresses the problem of mitigation of toxicity in Language models (LMs) for English language, even though there more than 7000 languages globally (Joshi et al., 2020) and future works should address more generalizable and multilingual solutions so that safety is promised for diverse set of speakers and not limited to English speakers (Weidinger et al., 2022)

**2. No one fixed toxicity taxonomy:** Literature survey highlights the fact that toxicity, hate and abuse and other related concepts are loosely defined and vary based on demographics and different social groups (Paz et al., 2020; Yin and Zubiaga, 2021). Henceforth, affecting the quality of hatespeech detection systems (HATEBERT) used in this work. These variations differences between cultural definitions of toxicity poses an ethical challenge (Jacobs and Wallach, 2021; Welbl et al., 2021).

**3. Third party classifiers for toxicity detection:** Reliance on the third party classifiers for toxicity detection can itself beat the purpose of fairness as these systems are reported to be biased towards certain protected groups and overestimate the prevelence of toxicity associated with them in the texts (Davidson et al., 2019; Abid et al., 2021; Hutchinson et al., 2020; Dixon et al., 2018; Sap et al., 2019a). For most part, we take care of these by using causal mechanisms but the ATE computation still involves using a toxicity classifier (HATEBERT) model.

## 10 Potential Risks

Any controlled generation method runs the runs the risk of being reverse-engineered, and this becomes even more crucial for detoxification techniques. In order to amplify their ideologies, extremists or terrorist groups could potentially subvert these models by prompting them to generate extremist, offensive and hateful content (McGuffie and Newhouse, 2020).

## 11 References

## References


Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. 2022. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv preprint arXiv:2202.11176*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2105.03023*.

Rahul Madhavan, Rishabh Garg, Kahini Wadhawan, and Sameep Mehta. 2023. CFL: Causally fair language models through token-level attribute controlled generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11344–11358, Toronto, Canada. Association for Computational Linguistics.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Debiasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*.

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Jose Quiroga Perez, Thanasis Daradoumis, and Joan Manuel Marques Puig. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, 28(6):1549–1565.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

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

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019b. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Mitigating spurious correlation in natural language understanding with counterfactual inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11308–11321.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

## A Proof of Theorem 1

**Theorem.** Given Assumptions 1 and 2, for $w_i$ which is a spurious correlate, $\texttt{ATE}(w_i) \leq 0.25$.

*Proof.* If we consider three numbers $\{\widehat{A}(c_i), \widehat{a}(w_i), \widehat{a}(w'_i)\}$, there are six possible orderings of this set. We can subsume these orderings into two cases:

1. $\widehat{A}(c_i) < \widehat{a}(w'_i)$.
2. $\widehat{A}(c_i) \geq \widehat{a}(w'_i)$.

Within these cases, we study the variation of $ATE(w_i)$ with $\widehat{a}(w_i)$. We plot these results in the Figure 5.

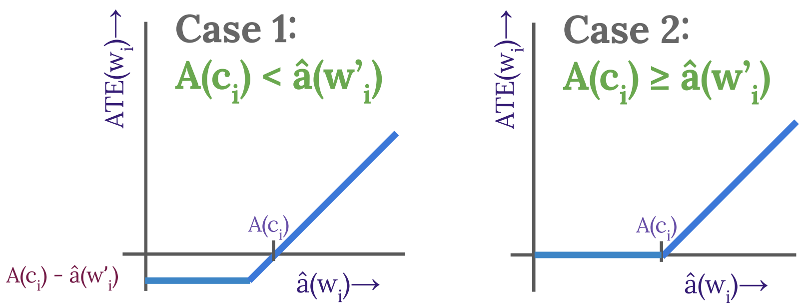


Figure 5: Graph of ATE score of a given word $w_i$ with $\widehat{a}(w_i)$ given two cases

Note that by Assumption 1, we have $\widehat{a}(w'_i) \leq \widehat{A}(c_i)$. Therefore, Case (2) in Figure 5 is sufficient for proof. We have:

$$\texttt{ATE}(w_i) = \mathop{\mathbb{E}}_{s \in \mathcal{D}} \mathop{\mathbb{E}}_{w'_i \in s'} \left[\widehat{A}(s) - \widehat{A}(s')\right] \tag{8}$$

$$\begin{aligned} &= \frac{n(A(s) = 1 \mid w_i \in s)}{n(s \mid w_i \in s)} \mathop{\mathbb{E}}_{w'_i \in s'} \left[\widehat{A}(s) - \widehat{A}(s')\right] \\ &+ \frac{n(A(s) = 0 \mid w_i \in s)}{n(s \mid w_i \in s)} \mathop{\mathbb{E}}_{w'_i \in s'} [\widehat{A}(s) - \widehat{A}(s')] \end{aligned} \tag{9}$$

But by Assumption 2, in toxic sentences, $\widehat{A}(s) = \widehat{A}(c_i) \geq \widehat{a}(w'_i)$. Therefore $\mathbb{E}_{w'_i \in s'}\{\widehat{A}(s) - \widehat{A}(s')\} = 0$. Then:

$$\texttt{ATE}(w_i) = \frac{n(A(s) = 0 \mid w_i \in s)}{n(s \mid w_i \in s)} \mathop{\mathbb{E}}_{w'_i \in s'} [\widehat{A}(s) - \widehat{A}(s')] \tag{10}$$

But $\widehat{A}(s) - \widehat{A}(s')$ is at most $\widehat{a}(w_i)$ as:
(1) if $\widehat{a}(w_i) \leq \widehat{A}(c_i)$, then $\widehat{A}(s) - \widehat{A}(s') = 0$
(2) otherwise $\widehat{A}(s) - \widehat{A}(s') = \widehat{a}(w_i) - \widehat{A}(s') \leq \widehat{a}(w_i)$. Then:

$$\texttt{ATE}(w_i) \leq \frac{n(A(s) = 0 \mid w_i \in s)}{n(s \mid w_i \in s)} \widehat{a}(w_i) \tag{11}$$

$$\begin{aligned} &= \frac{n(A(s) = 0 \mid w_i \in s)}{n(s \mid w_i \in s)} \frac{n(A(s) = 1 \mid w_i \in s)}{n(s \mid w_i \in s)} \\ &= p \cdot (1 - p) \end{aligned} \tag{12}$$

for some $p \in [0, 1]$. But $p \cdot (1 - p) \leq 0.25 \ \ \forall p \in [0, 1]$. □

Based on Theorem A and Lemma 1, $A(s) \leq 0.25$ if each $w_i \in s$ is a spurious correlate, i.e. non-causal, for attribute $A$.

# B Experimental Results in Detail for Zampieri et al. and Gao et al. Datasets

In this section we provide the full set of results on our runs across models for the two datasets Gao and Huang (2017) and Zampieri et al. (2019a). The plot in 6 illustrates the reduction in toxicity classification by using ATE score on the Zampieri et al. (2019a) dataset for three types of classifiers.
We provide the full tabular results in Tables 4 and 5.

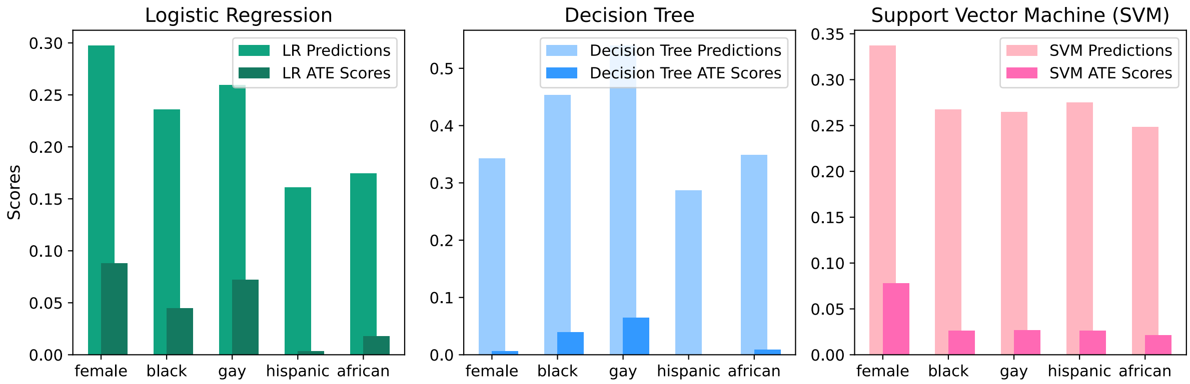


Figure 6: For the Zampieri et al. (2019a) dataset, we compute the mitigation of toxicity score using three different classifiers, and the ATE scores computed using the respective classifiers. These show a reduction on toxicity for protected groups across different models.

Table 4: Classifier Metrics by Protected Category for the Gao et al. Dataset

| Group → | African | | | Black | | | Female | | | Gay | | | Hispanic | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model ↓ | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff |
| LR | 0.201 | 0.099 | 0.102 | 0.300 | 0.108 | 0.192 | 0.270 | 0.167 | 0.103 | 0.470 | 0.167 | 0.303 | 0.166 | 0.011 | 0.155 |
| SVM | 0.282 | 0.062 | 0.220 | 0.282 | 0.052 | 0.230 | 0.301 | 0.082 | 0.219 | 0.371 | 0.154 | 0.217 | 0.246 | 0.057 | 0.189 |
| GB | 0.225 | 0.052 | 0.173 | 0.335 | 0.071 | 0.264 | 0.225 | 0.000 | 0.225 | 0.653 | 0.204 | 0.449 | 0.225 | 0.020 | 0.205 |
| NB | 0.460 | 0.002 | 0.458 | 0.510 | 0.047 | 0.463 | 0.444 | 0.004 | 0.440 | 0.657 | 0.107 | 0.550 | 0.615 | 0.000 | 0.615 |
| NN1Layer | 0.000 | 0.003 | -0.003 | 0.000 | 0.059 | -0.059 | 0.000 | 0.024 | -0.024 | 1.000 | 0.197 | 0.803 | 0.000 | 0.000 | 0.000 |
| NN2Layer | 0.000 | 0.000 | 0.000 | 0.000 | 0.096 | -0.096 | 0.002 | 0.000 | 0.002 | 1.000 | 0.217 | 0.783 | 0.000 | 0.000 | 0.000 |
| NN3Layer | 0.000 | 0.160 | -0.160 | 0.000 | 0.097 | -0.097 | 0.000 | 0.000 | 0.000 | 0.993 | 0.165 | 0.828 | 0.000 | 0.000 | 0.000 |

Table 5: Classifier Metrics by Protected Category for the Zampieri et al. Dataset

| Group → | African | | | Black | | | Female | | | Gay | | | Hispanic | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model ↓ | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff | Pred | ATE | Diff |
| LR | 0.174 | 0.020 | 0.154 | 0.236 | 0.049 | 0.187 | 0.297 | 0.075 | 0.223 | 0.260 | 0.098 | 0.162 | 0.161 | 0.143 | 0.018 |
| SVM | 0.248 | 0.030 | 0.218 | 0.267 | 0.036 | 0.232 | 0.337 | 0.068 | 0.269 | 0.265 | 0.033 | 0.232 | 0.275 | 0.119 | 0.156 |
| GB | 0.269 | 0.020 | 0.249 | 0.269 | 0.013 | 0.256 | 0.269 | 0.008 | 0.261 | 0.269 | 0.003 | 0.266 | 0.269 | 0.033 | 0.236 |
| NB | 0.349 | 0.009 | 0.341 | 0.453 | 0.055 | 0.398 | 0.343 | 0.183 | 0.160 | 0.539 | 0.070 | 0.469 | 0.287 | 0.000 | 0.287 |
| NN1Layer5 | 0.000 | 0.000 | -0.000 | 0.000 | 0.052 | -0.052 | 0.000 | 0.000 | -0.000 | 0.000 | 0.114 | -0.114 | 0.000 | 0.000 | 0.000 |
| NN2Layer105 | 0.000 | 0.000 | 0.000 | 0.000 | 0.090 | -0.090 | 0.000 | 0.170 | -0.170 | 0.000 | 0.104 | -0.104 | 0.000 | 0.000 | 0.000 |
| NN3Layer20105 | 0.000 | 0.200 | -0.200 | 0.000 | 0.126 | -0.126 | 0.000 | 0.075 | -0.075 | 0.000 | 0.046 | -0.046 | 0.000 | 0.000 | 0.000 |

**Note:** We note that the neural classifiers may have overfit on the Zampieri et al. (2019a) dataset due to which the numbers are either close to 0 or 1.

# C Experimental Setup

## C.1 Dataset Details

We conducted experiments on the publically available Zampieri (Zampieri et al., 2019b) and Gao (Gao and Huang, 2017) datasets.

## C.2 Hyper-parameters

Details in our GitHub repository: github.com/causalate-mitigates-bias/causal-ate-mitigates-bias

## C.3 Result Statistics

Our run details are provided on the README.md file of our GitHub repository: https://github.com/causalate-mitigates-bias/causal-ate-mitigates-bias/blob/main/README.md

## C.4 Compute Resources

All our experiments were carried out using NVidia 1080 GPU Machines with Intel Core i7-7700K @ 4.2GHz. Our experiments utilized approximately 100 CPU-hours and 10 GPU-hours.

## C.5 Tools and packages

We list the tools used in our requirements.txt file of our GitHub repository: https://github.com/causalate-mitigates-bias/causal-ate-mitigates-bias/blob/main/requirements.txt

## C.6 Use of AI Assistants

We have used AI Assistants (GPT-4) to help format our charts as well as help create latex tables.