# Bias detection and reduction in toxic langauge systems

Student Name: M. Gangaidzo
Supervisor Name: S. Concannon
Submitted as part of the degree of BSc Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

**Abstract**—Language models, similar to machine learning models, can display discriminatory behaviors due to statistical bias. This bias can arise from the models being trained on data that reflects real-world asymmetries and biases, causing the models to oversimplify complex relationships between variables to optimize accuracy. This leads to potential unfair treatment of individuals solely based on shared characteristics with others in real-world situations with situational asymmetrical population distributions. The paper focuses on the impact of statistical bias in a BERT model adapted for toxic language classification and proposes a debiasing technique using counterfactual data augmentation to demonstrate its effectiveness. The related work section highlights the efforts made to enable this type of research as well as other debiasing methods that have been used in similar tasks to debias language models. The study finds that the proposed methods can reduce bias in toxic language systems, but there is still room for improvement. Ultimately, the paper emphasizes the need for continued research and development in this area to ensure ethical and effective deployment of toxic language systems.

**Index Terms**—Language models, Machine learning, Natural Language Processing, Text Classification, Text processing

## 1 Introduction

The rise of media and online platforms has transformed communication, enabling people to connect across the globe and share information, build communities, and network more easily. However, these platforms have also made it easier for users to spread messages of hate while remaining anonymous, underscoring the need for effective toxic language detection systems (TLDS). While these systems can help detect and remove offensive and hateful content, they may also suffer from statistical biases that can result in discriminatory outcomes.

TLDS often rely on discriminative models, which can introduce biases during training due to skewed data populations or labeling biases. Biases can arise during data collection or labeling, and can also be amplified by model architecture as models optimize their loss functions based on asymmetrical data distributions. Biased TLDS can result in unfair censorship or a lack of censorship when needed, making it essential to develop debiased models.

One promising method for reducing direct gender bias in language and occupational associations is Counterfactual Data Augmentation (CDA). While this technique has not been used extensively for debiasing language models trained on toxic language classification, it holds potential for addressing bias in these systems.

To test the effectiveness of CDA on toxic language detection, we implemented a Bidirectional Encoder Representations from Transformers (BERT) model from Huggingface and trained it on both CDA and non-CDA toxic language classification datasets from Davidson et al. (2017). We evaluated the model for gender and racial bias markers as well as accuracy, using the Word Embedding Association Test (WEAT) to measure bias before and after training. Our experiments revealed that the dataset did not induce association bias, but we investigated whether any bias existed towards African American English (AAE) at similar levels to those demonstrated by Zhou et al. (2021).

## 2 Related Work

Modern toxic langauge detection systems employ different types of ML langage models in order to acheive their function. As previously stated, these models can be biased which can result in unfair behaviour. In this section I introduce the language model "BERT", which was used as the ML model for the toxic langauge classification system implementation and work previosly done to reduce bias present in BERT while preserving performance. I then discuss relevant work done to measure the bias present in language classification systems using the Word Embedding Association Test and the effectivness of CDA, DR and INLP, demonstrated in previous works, in the debiasing of language classification systems.

### 2.1 BERT

The BERT model was first introduced in 2019 by Devlin et al. (2019) in their research paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" . BERT is a model that is designed to pre-train deep bidirectional representations from unlabeled data by jointly conditioning on both left and right context in all layers. The pre-trained BERT model can be fine-tuned and adapted to a wide range of tasks such as question answering or text classification, simply by adding one additional output layer. With this simple adjustment, BERT is able to provide state-of-the-art performance in most NLU tasks and remains the point of reference in most cases.

#### 2.1.1 Pre-training

Pre-trained word embeddings refer to word embeddings that have been learned in other tasks and are subsequently utilized in similar tasks. These word embeddings have become a crucial component of training natural language processing (NLP) systems, as they offer significant enhancements in comparison to word embeddings that are learned from scratch, according to Turian et al. (2010). The use of pre-trained word embeddings not only provides the model with additional contexts with which words are used, which subsequently increases the accuracy of associations established in subsequent training, but also diminishes the likelihood that the model learns relationships that are exclusively present in the dataset of the current task. Erhan et al. (2010) observed that their research results indicated that unsupervised pre-training facilitates better generalization by directing learning towards better weights and biases that minimize the margin of error in predictions. Research into pre-training has continuously demonstrated that it provides substantially significant boost in model performance. However there is a risk of biases from the pre-training dataset being induced in the language model. As stated by Levine et al. (2022), although modern natural language modelling techniques have provided a significant boost in the natural language processing landscape, it has exposed the need for these machine learning models to be more nuanced in their understanding of text. Whether this will require improvements to be made in the pre-training process or model architectures is an active field of research.

#### 2.1.2 BERT in text classification

In 2021 Gonzalez-Carvajal et al. published a paper in which they demonstrated BERT's ability to perform sentiment analysis in movie reiviews. The aim of their research was to compare BERT's performance to traditional machine learning methodologies. Where BERT is able to adapt well most language understanding tasks, traditional machine leanring methods require the best model to be found for each context which requires time and resorces. BERT achieved a sentiment classification accuracy of 93.87% which out performed 6 other machine learning models by an average margin of 3%. They then performed another text classification experiment in which tweets regarding natural disasters were classed as real or fake. In this experiment their model was able to achieve a high accuracy score of 83.61% which outperformed the H20AutoML model by a margin of 6%. It is possible that data processing and model tuning would have been able to reduce the gap between BERT and the machine learning models. However, this shows the main reason why BERT has become high favoured for many NLU tasks. With relatively little effort and time, it is easier to obtain close to state-of-the-art results with BERT. It's versitility and capability in natural language understanding tasks substantially higher than most traditional machine learning architectures.

#### 2.1.3 Model Architecture

The architecture proposed by Devlin et al. (2019) was a multi-layer bidirectional transformer encoder. This architecture was based on an original implementation of transformers by Vaswani et al. (2017) in their paper "Attention is all you need" which is shown below in figure 1. The Transformers architecture uses an encoder to map an input sequence of symbol representations ($x_1$, ...., $x_n$) to a sequence of continuous representations z = ($z_1$, .... , $z_n$ ). Given an input $z$, the decoder then generates an output sequence of symbols one element at a time. The decoder takes the output of the encoder and generates the output sequence, also using self-attention and feed-forward neural networks. The model is auto-regressive and takes the previously generated symbols as additional input when generating the next. Transformers have this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder (Vaswani et al. 2017). Transformers have become a popular model in natural language processing (NLP) tasks such as language translation, language understanding, and text generation. The use of self-attention mechanisms allows the model to selectively focus on different parts of the input sequence when making predictions. This is different to traditional sequence-to-sequence models that use recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to process the input sequentially. During training, the model learns to predict the correct output sequence given the input sequence by minimizing a loss function. During inference, the model generates the output sequence one token at a time by recursively predicting the next token based on the previous ones.

Overall, the transformers architecture has shown state-of-the-art performance in many NLP tasks and has become a cornerstone of modern NLP research and development and has allowed for the development of the BERT model discussed in this paper.

In the paper by Devlin et al. (2019) they detail two model sizes $BERT_{BASE}$ , where is has 12 layers, 768 hidden size and 12 attention heads with a total of 100 million parameters, and $BERT_{LARGE}$, where it has 24 layers, a hidden size of 1024, and 16 attention heads with a total of 340 million parameters. The base model had the same number of parameters as Open AI's GPT model. BERT takes input embeddings which are the sum of the token embeddings, the segmentation embeddings and the position embeddings. as shown in figure 2

### 2.2 Word Embedding Association Test

The Oxford dictionary definition of bias is, the prejudice for or against an individual or group in a manner that can be considered unfair. There are a number of ways to measure bias in classification algorithms. One of these ways is the Word Embedding Association Test (WEAT) proposed by Caliskan et al. (2017). In their paper their results indicate that there are recoverable and accurate signs of historical biases exhibited in our langauge. These biases range from morally neutral topics such as instects and flowers to more problematic topics such as race and gender. In addition to these findings they also contribute the WEAT as a new way of evaluating bias in text. This new evaluation method was based on the Implicit Association Test (IAT) proposed by Greenwald et al. (1998) applied to the semantic representation of words in AI (word embeddings). This evaluation metric allowed for the evaluation of bias in word embeddings which are widely used in the state-of-the-art language models.

Fig. 1. Transformers architecture

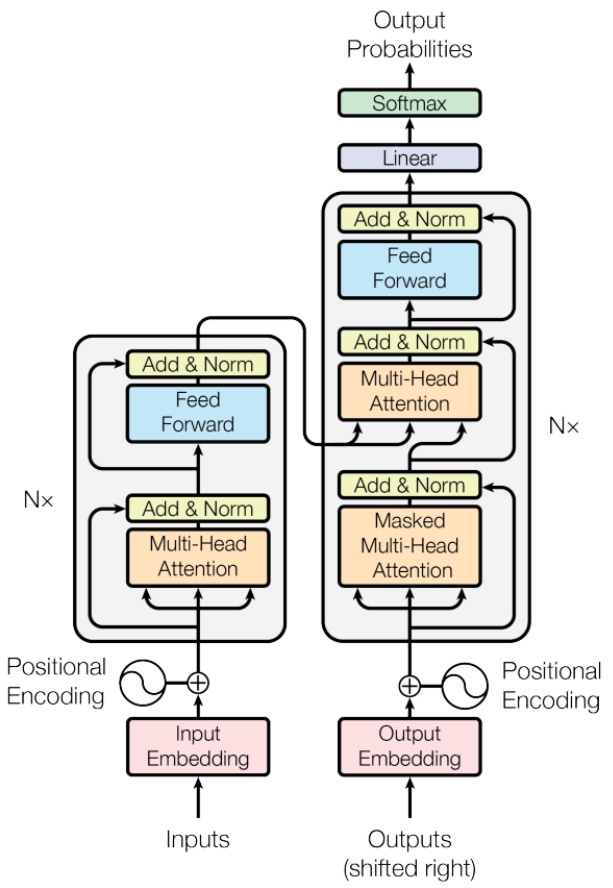
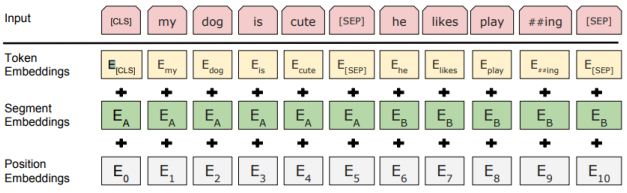



Fig. 2. BERT input representation




## 2.3 Word Embedding debiasing

Word embedding debiasing methods are employed to mitigate the influence of the biases that can be acquired during word embedding training. Since word embeddings represent semantic meanings, they can inherit the prejudices present in the training data, which can lead to problematic and unfair results when used in natural language processing (NLP) applications. There are several techniques proposed to address this issue, which can be broadly categorized as post-processing or pre-processing methods.

Post-processing methods, such as the Iterative Nullspace Projection (INLP) technique proposed by Ravfogel et al., modify the existing word embeddings to eliminate the bias. This approach aims to identify and project the embedding vectors that contribute to the bias onto a subspace orthogonal to the space spanned by the non-biased embedding vectors. The iterative projection process continues until the embeddings no longer contain biased information.

On the other hand, pre-processing methods aim to remove the bias present in the dataset or model architecture that will be used in training. This involves techniques such as re-sampling the training dataset to ensure equal representation of all groups, modifying the model architecture to reduce the impact of certain features, or even training new word embeddings on a debiased dataset and modified model. These methods can help ensure that the word embeddings learned do not encode any biased information.

In recent years, many successful debiasing methods have been proposed, such as Hard Debias by Bolukbasi et al., which is a pre-processing technique that modifies the training data to achieve a more balanced representation of gender and ethnicity. Another technique is the Adversarial Debiasing method proposed by Zhang et al., which uses a generative adversarial network (GAN) to generate a new embedding space that is free from biased information.

Overall, these methods have shown promise in reducing bias in word embeddings, which can have significant implications in real-world NLP applications, particularly in areas such as hiring, lending, and sentiment analysis, where biased word embeddings can lead to unfair and discriminatory outcomes.

### *2.3.1 Itterative Nullspace Projection*

The Iterative Nullspace Projection (INLP) technique has demonstrated its effectiveness in reducing the linear separability between female-biased and male-biased words, thereby improving the quality of word embeddings. According to Ravfogel et al. (2020), the technique increases the correlation between word embeddings and human judgment, indicating that it reduces the influence of biased information in the training data. The researchers also used the Word Embedding Association Test (WEAT) to evaluate the extent to which INLP addresses the "bias-by-neighbours" phenomenon, which refers to a type of bias that can occur in word embedding models. This phenomenon arises when the embeddings of words become biased due to their frequent co-occurrence with other words in the training data, leading to the embeddings of words becoming associated with stereotypical attributes.

For instance, Bolukbasi et al. (2016) demonstrated this phenomenon by showing that the embeddings developed an association between male/female words and traditional male/female stereotypes, respectively, due to the context in which they were trained. While other debiasing methods are more effective at reducing direct bias, they fail to address indirect bias. In contrast, INLP reduces the impact of indirect bias by projecting protected attributes into the nullspace. The technique shows promise in reducing the biased association between unrelated words (Ravfogel et al. 2020), indicating its ability to reduce the influence of indirect bias in word embeddings.

## 2.4 Counterfactual Data Augmentation

Counterfactual data augmentation is a methodology for corpus augmentation. It is a simple intervention strategy that results in the breaking of association between the concerned subjects by inverting the sentences bias in the

other direction and appending the inverted sentence to the training dataset. This approach attempts to equalise the relationship a set of attributed words has with neutral words (i.e. words that should not be more strongly correlated with any of the words in the attributed word set). Successfully implementing this intervention, results in a bias-reduced corpus. An easy example of this intervention would be, ”He is a business owner and entrepreneur”, would have a counter example created to have ”She is a business owner and entrepreneur”. between a set of attribute words and a set of target words by balancing the the number of times that the set of attribute words are mentioned in the corpus. by mentions in the dataset Lu and colleagues used this to mitigate the gender bias in their corpus. This process would happen for all attribute mentions in the corpus and then the embedding would be trained on the augmented corpus. If done properly, this should mitigate the he-she bias in the corpus. Lu et al.'s improved CDA intervention does not swap instances where the subject is corefered to by a proper noun. This is done to avoid creating sentences with unintended associations. For example, we would not necessarily want a sentence which has the words ”John... he.... groom”, to be changed to, ”John... she ... bride” . Not only does this potential to create confussing word associations, but it could result in reduced performance for any models that use the word embedding since this could produce associations between words that should not be associated or it could reduce associations between words that are closely associated by increasing the statistical chance of other words occurring in those contexts. CDA is a rather simple method to implement and low in computational cost (at least compared to other types of debiasing methods). However, the disadvantage of CDA is that it will increase the size of the training dataset by a factor of $n$ (in the best case) where n is the number of attributes being augmented. Furthermore, running CDA becomes more computationally expensive depending on how many attributes you are changing. Consider changing 1 type of attributes (gender (male, female)) and changing 3 types (gender, ethnicity, religion). When changing gender, it is a lot quicker to compute but when changing ethnicity and religion, the list of swaps very quickly increases which may result in an extremely large dataset. In the context of running experiments this may not be a problem since we can limit the size of the concerned sets. However, in real world applications, all groups would have to be included in order to implement this debiasing method fairly. To take it a step further, suppose a bias is discovered towards nationality, implementing CDA on a nationality level would not be feasible to implement to allow for all countries to have a CDA instance because the number of CDA instances exponentially with the number of mentions of countries. This is to say, CDA similar to other debiasing methods has its advantages and disadvantages and is not applicable in all cases.

### 2.5 CDA Alternative (CDS)

CDS or Counterfactual Data Substitution, is a technique proposed by Maudslay et al. They propose this mehod because conventional CDA produces debiased corpa with what they call ”perculiar statistical properties”. Measurements such as word frequencies being even (because the corpus is doubled) and type-token ratio is observed to be lower than predicted by Heaps' Law (a law that states the vocab size of a corpus grows as a power-law function of the corpus size). Maudslay note that the effect this can have on the embedding is not so easy to predict and they make the assumption it is better to avoid violating this law if possible. Instead, they apply substitutions (with a 0.5 probability) which produces a non-duplicated counterfactual training corpus. This is the techniques they call Counterfactual Data Substitution. The results were that the CDS implementation outperformed CDA in their implementation. It successfully reduced direct gender bias in 80% of the test cases.

## 3 Methodology

In this investigation, I examined the effectiveness the debiasing method, ”Counterfactual Data Augmentation” (CDA) in the context of toxic language classification using the BERT model. CDA is a technique used to reduce bias in machine learning models by generating new data points that are similar to the original data points, but with some substitutions made to create identical instances of specified attributes such as gender or race in the hopes that this will balance the identity mentions in the dataset thereby reducing bias.

To test the effectiveness of CDA, I conducted experiments where I compared the levels of bias in six different instances of the BERT model. I itteratively changed a numbder of parameters which affected whether the model had undergone pretraining or training and whether the data it was trained on had undergone CDA. By comparing the bias levels across these instances, I aimed to determine the impact of CDA on reducing bias in the BERT model.

Overall, this investigation aimed to provide insight into the effectiveness of CDA as a debiasing method in the context of toxic language classification using BERT. This is an important research area, as biased language models can perpetuate harmful stereotypes and prejudices, particularly in applications such as social media moderation and online content filtering. The results could have important implications for the development of more fair and unbiased language models in the future.

### 3.1 Dataset

In this investigation, the dataset provided by Davidson et al. (2017) for automated hate speech detection was utilized. The dataset was created by randomly sampling 8.5 million tweets obtained from Harebase.org and selecting around 25,000 tweets that contained terms from a hate speech lexicon. These tweets were then labeled using CrowdFlower workers, who were instructed to consider not only the words used but also the context in which they appeared when deciding whether a tweet was part of the offensive, hate speech, or neither category. At the start of the research, the dataset consisted of 24,783 labeled comments for training and validation, with each comment being allocated one of three possible labels: Hate Speech, Offensive, or Neither. The label with the most votes was then set as the official category for the comment. However, the researchers found that tweets containing sexist content were more likely to

be identified as offensive, while tweets containing racist or homophobic content were more likely to be classified as hate speech. This suggests that there were inherent biases present in the individuals who aided in labeling the dataset. A shortened version of the dataset is shown in the table below in table 1. The columns that contained the number of votes as to the labe as well as the number of people who were involved in labeling the comments were dropped. There is a possibility this information could have been useful as it may have helped the model understand how likely the label allocated was the correct label. However, due to time contraints and performance not being the primary focus of this research (with the primary focus being the bias induced), I elected to drop these columns and instead change this to a classification problem of 3 categories where the model had to predict which of the 3 categories the comments belonged to.

TABLE 1
Dataset containing comments and the relevant labels

| | Unnamed: 0 | count | hate_speech | offensive_language | neither | class | tweet |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 3 | 0 | 0 | 3 | 2 | !!! RT @mayasolovely: As a woman you shouldn't... |
| 1 | 1 | 3 | 0 | 3 | 0 | 1 | !!!!! RT @mleew17: boy dats cold...tyga dwn ba... |
| 2 | 2 | 3 | 0 | 3 | 0 | 1 | !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby... |
| 3 | 3 | 3 | 0 | 2 | 1 | 1 | !!!!!!!! RT @C_G_Anderson: @viva_based she lo... |
| 4 | 4 | 6 | 0 | 6 | 0 | 1 | !!!!!!!!!!! RT @ShenikaRoberts: The shit you... |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 24778 | 25291 | 3 | 0 | 2 | 1 | 1 | you's a muthaf***in lie “@LifeAsKing: @2... |
| 24779 | 25292 | 3 | 0 | 1 | 2 | 2 | you've gone and broke the wrong heart baby, an... |
| 24780 | 25294 | 3 | 0 | 3 | 0 | 1 | young buck wanna eat!!.. dat nigguh like I ain... |
| 24781 | 25295 | 6 | 0 | 6 | 0 | 1 | youu got wild bitches tellin you lies |
| 24782 | 25296 | 3 | 0 | 0 | 3 | 2 | ~~Ruffled \| Ntac Eileen Dahlia - Beautiful col... |

I took a closer at the dataset to see what words most commonly occured when the offensive or hateful labels were allocated.

### 3.2 Training

The aim of this study was to investigate the extent to which bias was induced in the BERT model by the toxic language dataset, as well as the bias that was induced by pretraining the model with the Huggingface implementation. To achieve this, several parameters were modified, such as whether the dataset underwent Counterfactual Data Augmentation (CDA) and whether or not the model was pre-trained. By measuring the bias induced in the model at different stages, we were able to observe the types of biases that were induced with respect to the training data.

To begin the study, the Huggingface implementation of the BERT model was randomly initialized. It was also necessary to ensure that any biases present in the model were not solely due to pretraining performed by Huggingface. While it is acknowledged that the power of the BERT model lies in its pretraining performance, it was essential to control for this factor to measure the impact of debiasing techniques on both the bias and performance of the model. By controlling for the pretraining factor, we were able to isolate the impact of training on the toxic language dataset on model bias and performance.

The CDA technique was then applied to the toxic language dataset to generate additional data points that were similar to the original dataset but with some key differences that helped reduce bias. This augmented dataset was then used to train the BERT model, and the bias was measured once again. By comparing the bias levels between the augmented and original datasets, we could evaluate the effectiveness of CDA in reducing bias in the BERT model.

During the training spilts the goal of identifying the extent of the bias induced in the BERT model by both the toxic language dataset and pretraining was a primary focus. By controlling for pretraining and utilizing debiasing techniques such as CDA, we aimed to develop a more fair and unbiased language model. The results of this study have important implications for the development of machine learning models that can be deployed in sensitive contexts, such as online content filtering and social media moderation.

For this study, we obtained a manually labeled dataset of tweets from Davidson et al. The tweets were labeled according to the number of individuals who contributed to the labeling process, the number of individuals who designated the tweet as belonging to label A, B, or C, and the overall class of the tweet. The labels used in the dataset were A, B, C, D, and E, representing count, hate speech, offensive, neither, and class, respectively.

To illustrate, an example of the labeling process is shown below in table 2:

TABLE 2
Example of tweet labeling process

| Number of Labelers | A | B | C | Overall Class |
|---|---|---|---|---|
| 3 | 0 | 2 | 1 | B (Offensive) |

In this example, the tweet was labeled by three individuals. Two individuals designated the tweet as belonging to label B (Offensive), while one individual designated the tweet as belonging to label C. The overall class of the tweet was therefore designated as B. This labeling scheme was used to accurately categorize the tweets in the dataset and facilitate the analysis of the effectiveness of the debiasing method in reducing bias in the BERT model trained on this dataset.

### 3.3 Word Embedding Association Test

In this paper, we utilize the Word Embedding Association Test (WEAT), a method proposed by Caliskan et al., to quantify the bias in the BERT model by examining the bias embedded within its word embeddings. This technique is useful because it enables us to understand how the model perceives the relationship between a group of target words and a group of attribute words. For example, in the context of gender, an unbiased word embedding would exhibit gendered words with similar measured similarity when compared to a set of neutral words. This would be indicated by the WEAT outputting a score of 0, where values closer to 0 indicate low levels of bias, and values closer to 1 or -1 demonstrate bias in the direction of one or the other embedding. Other methods to measure bias are primarily focused on determining the model's tendency to generate an arbitrary output for an arbitrary input.

In order to measure the bias present in the BERT model, we use a technique called the Word Embedding Association

Test (WEAT). This test was first introduced in a paper by Caliskan et al. and aims to identify the degree of association between two sets of words. In our case, we are interested in measuring the association between gendered words (e.g. ”male”, ”man”, ”boy”, ”female”, ”woman”, ”girl”, etc.) and a set of target words related to career and family (e.g. ”executive”, ”management”, ”professional”, ”home”, ”parents”, ”children”, etc.). We want to ensure that the association between gendered words and the target words is balanced and not biased towards one gender or the other. To do this, we compare the association between the gendered words and the target words by calculating a statistical score. The null hypothesis is that there is no difference between the two sets of target words with respect to their similarity to the two sets of attribute words. By measuring this score, we can quantify the degree of bias present in the BERT model and determine the effectiveness of our debiasing techniques. This can be done not only for gender but other attributes of the same category. In formal terms we do the following:

- Choose two sets of attribute concepts, A and B. These sets can represent any binary categorization (e.g., male/female, African-American/European-American or pleasant/unpleasant)
- Choose two sets of words, X and Y, which are expected to be associated with A or B respectively.
- Calculate the Cosine Similarity between each word set X and Y, and each attribute concept in set A or B respectively.
- Compute the mean similarity scores of each set of words by averaging the calculated cosine similarities across all target concepts in the coresponding set.
- Lastly, we compute the difference between the mean similarity scores for sets A and B, and compare it to the difference between the mean similarity scores for sets X and Y.
- The test statistic (taken from Caliskan et al. 2017) is:

$$s(X, Y, A, B) = \sum_{x \epsilon X} s(x, A, B) - \sum_{y \epsilon Y} s(y, A, B) \quad (1)$$

where

$$s(w, A, B) = mean_{a \epsilon A} cos(\vec{w}, \vec{a}) - mean_{b \epsilon B} cos(\vec{w}, \vec{b}) \quad (2)$$

Therefore, it can be said that s(w, A, B) measures the association of word w with the attribute and the whole function s(X, Y, A, B) measures the difference in the association between the two sets of target words with respect to the attributes.

In order to evaluate the bias in the trained model I create a WEAT implementation based on the formula set out by Caliskan et al. This implementation calculates the WEAT score for attribute sets $A$ and $B$ for targets $X$ and $Y$, where $A$ and $B$ are two sets of different attributes that should show similar mean cosine similarity scores to target words from sets $X$ and $Y$. The p-value of these WEAT scores was also calculated in order to observe the probability of the obtained values.

### 3.4 Counterfactual data augmentation

In my investigation, I implemented the Counterfactual Data Augmentation (CDA) technique using attribute pairs for gender and race. The gender attribute pairs consisted of nine words each, with three words associated with the male attribute (”man”, ”male”, ”boy”) and six words associated with the female attribute (”woman”, ”female”, ”girl”). The African American and European American attribute lists had eight names each, which are commonly considered stereotypical of each group. Initially, there were longer lists of names, but since these names were not included in the pretraining of the BERT model, the model had no appropriate representation for them in its embeddings, rendering it unnecessary to look at the bias for such a case. I then went through each attribute list and created instances of the same sentence but with the opposite relevant attribute pair. This allowed me to observe how the model’s bias changed when it was trained on these counterfactual instances. By comparing the results from the original data and the augmented data, I was able to evaluate the effectiveness of CDA in reducing bias in the BERT model for toxic language classification.

## 4 Results

In this study, the results of the experiments conducted indicated that the bias present in the model regarding word associations was marginal, if present at all. There was no significant bias induced either by training or by the pre-training that was already present in terms of word association. However, upon further investigation, evidence of bias towards specific words (lexical bias) was found, similar to the findings of Zhou et al. (2021) in their research on challenges in automated debiasing for toxic language detection.

While the main focus of the study was on debiasing, it was noted that the accuracy scores of the models that had not undergone pre-training were significantly lower, with an accuracy rate of around 77%. These models also tended to fail in classifying sentences with words that they had seldom seen before, in comparison to the pre-trained models, which ranged in accuracy from around 90% to 95%. The pre-trained models only struggled with the appearance of specific words.

Overall, these findings suggest that pre-training plays a crucial role in the performance of the BERT model in toxic language classification, particularly in its ability to handle previously unseen words. While there is evidence of bias towards specific words, there is little to no bias in terms of word associations, indicating that the BERT model can be a reliable tool for toxic language classification. However, further research is necessary to determine the impact of this bias on the model’s overall performance and to develop effective debiasing methods.

### 4.1 Bias induced by training

The observable bias induced by training with the WEAT was found to be relatively small. Despite being trained on the hate speech dataset, the values did not change significantly. The study consisted of four training iterations, where the model was trained with and without pretraining before and after CDA. When comparing the trained model that did not undergo pretraining to the one that did, the results

varied. The WEAT test showed an increase in association between the female and African American attribute sets and their respective target word sets in the model that had undergone pretraining compared to the one that had not. However, the development of lexical bias was not observable or measured in the test results. Lexical bias refers to the bias shown towards specific words, which tends to result in the improper categorization of inputs in classification tasks. Zhou et al. (2021) demonstrated this problem in their paper and found lexical biases developing towards what they termed "African American English". In this task, due to the nature of the dataset, the developed model showed signs of the development of lexical biases towards words commonly used in offensive and/or hateful comments, including words that were not necessarily obscene. Providing the model with single word inputs occasionally caused it to provide the offensive tag when it was not appropriate. The model that showed the least amount of lexical bias was the one that had been trained on the regular dataset and had received pretraining before being trained on the downstream tasks. The results for the different experimental settings are presented in the four tables below.

TABLE 3
Bias Markers in non-pretrained BERT model trained on regular dataset

| | Attribute 1 | Attribute 2 | Target Words 1 | Target Words 2 | Weat Score a | Weat Score b | p_value |
|---|---|---|---|---|---|---|---|
| 0 | male attributes | female attributes | carrer words | family words | -0.008565 | -0.003711 | 0.630965 |
| 1 | male attributes | female attributes | math words | art words | 0.008295 | 0.001450 | 0.638220 |
| 2 | male attributes | female attributes | pleasant words | unpleasant words | 0.007399 | -0.008716 | 0.625192 |
| 3 | european american names | african american names | carrer words | family words | 0.000875 | -0.000451 | 0.661875 |
| 4 | european american names | african american names | math words | art words | 0.010220 | 0.006436 | 0.686604 |
| 5 | european american names | african american names | pleasant words | unpleasant words | -0.001371 | -0.015887 | 0.660635 |

TABLE 4
Bias Markers in pretrained BERT model trained on regular dataset

| | Attribute 1 | Attribute 2 | Target Words 1 | Target Words 2 | Weat Score a | Weat Score b | p_value |
|---|---|---|---|---|---|---|---|
| 0 | male attributes | female attributes | carrer words | family words | -0.001883 | -0.009666 | 0.797432 |
| 1 | male attributes | female attributes | math words | art words | -0.024864 | -0.025597 | 0.794221 |
| 2 | male attributes | female attributes | pleasant words | unpleasant words | 0.009511 | 0.005497 | 0.810678 |
| 3 | european american names | african american names | carrer words | family words | -0.004520 | -0.004497 | 0.831763 |
| 4 | european american names | african american names | math words | art words | -0.010545 | 0.006533 | 0.813634 |
| 5 | european american names | african american names | pleasant words | unpleasant words | -0.006040 | -0.015464 | 0.822904 |

## 4.2 Bias changes after CDA

In my investigation, I observed that implementing CDA led to a slight decrease in bias markers in the pretrained BERT model, particularly in the directions that were of concern. This decrease in bias was observed in most of the different biases measured, bringing the captured biases closer to 0. However, since the measured bias was already quite small, the amount by which it decreased was not significantly large.

It is my hypothesis that the reason for this limited decrease in bias may be due to the adjustments and tuning performed by Huggingface to mitigate racial and gender biases in their Python transformers models. However, it is also possible that the small decrease in bias observed was simply due to chance and unrelated to Huggingface's interventions.

To provide further insight into the impact of CDA, I have included tables containing the results for the experiments where CDA was implemented. Despite the marginal improvements observed in the WEAT score, the results suggest that further research is needed to determine the true effectiveness of CDA in reducing biases in pretrained language models.

TABLE 5
Bias Markers in non-pretrained BERT trained on CDA dataset

| | Attribute 1 | Attribute 2 | Target Words 1 | Target Words 2 | Weat Score a | Weat Score b | p_value |
|---|---|---|---|---|---|---|---|
| 0 | male attributes | female attributes | carrer words | family words | -0.008565 | -0.003711 | 0.630965 |
| 1 | male attributes | female attributes | math words | art words | 0.008295 | 0.001450 | 0.638220 |
| 2 | male attributes | female attributes | pleasant words | unpleasant words | 0.007399 | -0.008716 | 0.625192 |
| 3 | european american names | african american names | carrer words | family words | 0.000875 | -0.000451 | 0.661875 |
| 4 | european american names | african american names | math words | art words | 0.010220 | 0.006436 | 0.686604 |
| 5 | european american names | african american names | pleasant words | unpleasant words | -0.001371 | -0.015887 | 0.660635 |

TABLE 6
Bias Markers in pretrained BERT trained on CDA dataset

| | Attribute 1 | Attribute 2 | Target Words 1 | Target Words 2 | Weat Score a | Weat Score b | p_value |
|---|---|---|---|---|---|---|---|
| 0 | male attributes | female attributes | carrer words | family words | -0.001154 | -0.009663 | 0.794769 |
| 1 | male attributes | female attributes | math words | art words | -0.024770 | -0.025979 | 0.793270 |
| 2 | male attributes | female attributes | pleasant words | unpleasant words | 0.009450 | 0.005729 | 0.809636 |
| 3 | european american names | african american names | carrer words | family words | -0.004557 | -0.004663 | 0.832743 |
| 4 | european american names | african american names | math words | art words | -0.010568 | 0.006389 | 0.817166 |
| 5 | european american names | african american names | pleasant words | unpleasant words | -0.005805 | -0.015736 | 0.823245 |

### 4.2.1 *Lexical bias*

Lexical bias is a major issue in natural language processing, particularly in language modelling, where the output of the model can be significantly influenced by the presence of specific words or phrases without considering their context. This can happen due to the model's training data, where it may learn certain relationships between words or biases due to the skew in the distribution of specific words in the data. Additionally, biases can develop due to associations between words formed by society, such as the correlation between certain ethnicities and negative connotations.

In order to evaluate the effectiveness of the Counterfactual Data Augmentation (CDA) technique in reducing biases, several experiments were performed, and the results were documented. Following these experiments, I conducted input experiments to confirm the presence of lexical biases in the model. These experiments involved testing the model with a variety of inputs, varying the context of specific words, and observing the resulting output. Through this testing, I was able to identify areas of bias and work to mitigate them through further training and fine-tuning of the model. Since this had not initially been a point of focus I only performed experiments with the model with the highest accuracy

**Bias in identity mentions**: During the analysis of the performance of the model on the toxic language dataset, it was observed that the model exhibited different levels of bias when dealing with identity mentions of different types. Specifically, when identity mentions with regards to race were present in the text, the model was more likely to assign an offensive or hateful label as compared to when gender mentions were present. This observation is partly

explained by the biases present in the dataset due to the labeling process by the human labelers, as documented by Davidson et al. However, during the evaluation of the model's performance, some unexpected results were also observed. For example, when the model was prompted to classify the text "Women belong in the kitchen", it assigned a "Neither" tag, which should have been at least offensive if not hateful. A similar ambiguous case, "She belongs in the kitchen", which could be interpreted in different ways based on context, also resulted in the same "Neither" tag. This suggests that the model might struggle with ambiguity in the input and may not know which label to assign in such cases, as the two inputs are similar in the eyes of the model. These observations indicate the need for further research to improve the model's ability to handle identity mentions and deal with ambiguity in the input.

**Bias towards specific words**

During the evaluation of the BERT model, it became apparent that the model exhibited bias towards words that are frequently used in hateful and offensive contexts. For instance, when presented with sentences such as "Sam is a thug" or "Jamal is a thug," the model designated the "Hate" label. Although this sentence could be considered offensive, it highlights the ambiguity faced by classification systems. Even if a person was asked to classify this comment, the lack of context would make it difficult to determine if the comment is hateful or not. Additionally, the model demonstrated a bias towards obscene words, immediately categorizing sentences as "Offensive" if they contained an obscene word. However, it missed the "Hate" classification when presented with the input "all black people are dumb." This result is perplexing since the model did correctly identify the case where the sentence instead reads "black people are dumb," immediately classifying it as hateful if there is just the identity mention "black people" and no other text.

One potential upside is that the model did not designate different labels based on the name difference in the "Sam" and "Jamal" examples. This demonstrates that the model did not create a strong association between race and names in this specific case. Nevertheless, these examples demonstrate that the BERT model is still susceptible to lexical biases that exist due to the toxic language classification training. Unfortunately, I was unable to conduct more in-depth and rigorous research into the lexical biases since I had caught these examples late and did not have enough time to examine them further.

## 5 Evaluation

The project aimed to investigate the effectiveness of Counterfactual Data Augmentation (CDA) in debiasing toxic language detection systems. Throughout the course of the project, significant progress was made towards achieving this goal through the proposed interventions and implemented models. However, the results also revealed the need for further work to verify certain findings and the correctness of the implementation outlined in the paper.

The study demonstrated that the model was capable of fulfilling its task, but it still required debiasing to address its lexical biases towards certain words. CDA was found to have a slight impact on the model's performance in the pre-trained case, with a 1% drop in accuracy. However, it remains unclear if the change in accuracy was due to CDA or other factors. The research also revealed that a more in-depth implementation of CDA could have been used, and a thorough inspection of the dataset was necessary to determine the most appropriate way of implementing CDA. Additionally, the debiasing technique Counterfactual Data Substitution might have been a more suitable implementation, as CDA could potentially induce strange statistical phenomena within the data, thereby reducing the model's ability to observe relationships between words.

As I delved deeper into my research, I realized that the issue of lexical bias in text classification models was more complex than I initially anticipated. While I had focused on addressing direct racial and gender biases in the dataset, I had not fully considered how these biases might manifest through lexical bias. As I continued to explore the topic, I found that language and dialectical variation were often associated with certain groups and that these variations could be subject to prejudice and discrimination. This realization highlighted the need for a deeper understanding of how these biases can be addressed in text classification models.

In particular, I believe that it is crucial to find ways to reduce the impact of lexical biases towards certain ways of speaking, especially given the increasing role of social media in modern life. Such biases can perpetuate systemic discrimination and hinder effective communication between different groups. To this end, I plan to continue exploring ways to incorporate techniques such as Counterfactual data substitution to reduce lexical biases in future iterations of my model. By doing so, I hope to contribute to the development of more inclusive and equitable text classification models that can better serve diverse communities.

In retrospect, I realize that one way to improve the accuracy of the model could have been to increase the size of the dataset by incorporating data from multiple sources. While using a single dataset provides consistency to the classification system, it also poses a risk of biasing the model towards the specific language and patterns present in that dataset. By limiting myself to using only the dataset from Davidson et al. (2017), the model may have formed stronger associations between certain words and classification labels, which could have contributed to the lexical biases observed in the model.

Incorporating data from multiple sources could have provided a more diverse range of language patterns and usage, helping to reduce the model's reliance on specific words or phrases. Additionally, including data from sources with different demographics and backgrounds could have helped to mitigate any existing biases in the training data. However, using multiple sources can also introduce challenges in data cleaning and consistency, as well as the need to ensure that the datasets are properly balanced and representative.

Going forward, it will be important to consider the benefits and risks of using a single dataset versus multiple sources, and to carefully balance the need for consistency with the need for diversity and inclusivity in the training data. Ultimately, the goal should be to develop models that are both accurate and free from biases, in order to ensure fair and equitable outcomes in toxic language detection and other natural language processing applications.

Despite these challenges, the implementation of a toxic language classification system using BERT was successful. However, further research is needed to improve the system's debiasing capabilities and verify the findings. Future work could include exploring alternative debiasing techniques, conducting a more thorough analysis of the dataset, and examining the impact of different population skew in the data on the model's performance. Overall, this study represents a valuable contribution to the field of NLP, highlighting the need for more robust and fair models in detecting and combating hate speech and other forms of toxic language online.

In evaluating the results of the BERT model trained on the toxic language dataset, it is important to note that the model demonstrated clear biases towards certain words and contexts that are often associated with hate speech or offensive language. This was evident in the model's response to prompts such as "Sam is a thug" or "Jamal is a thug," which both resulted in the "Hate" label. While this may be considered an offensive sentence, it highlights the problem of ambiguity that classification systems face when it comes to determining the context and intent behind certain words or phrases.

Furthermore, the model showed a bias towards obscene words and immediately classified sentences as "offensive" if they contained such language, but it missed the "hate" classification for examples such as "all black people are dumb." It is worth noting, however, that the model correctly identified the hateful connotations of a sentence such as "black people are dumb," and immediately classified any comment as hateful if it contained the identity mention "black people" and no other text.

Overall, while the model demonstrated some biases and limitations in its classification capabilities, it also showed some promising results in correctly identifying hate speech in certain contexts. However, further research and analysis is needed to fully understand the extent of the lexical biases present in the toxic language dataset and how they impact the performance of the model. Additionally, more rigorous testing and evaluation methods should be employed to ensure that the model is not perpetuating harmful stereotypes or contributing to discriminatory practices.

## 6 Conclusion

In this paper, the harmful effects of bias in machine learning implementations have been discussed, specifically in the context of language models used for detecting and combating hate speech and other forms of toxic language online. It has been established that biases can lead to unjustifiable discrimination towards individuals or groups, which is a significant concern that needs to be addressed.

To mitigate bias in machine learning models, various methods have been proposed such as pre-training and post-training methods. Counterfactual data augmentation (CDA) is a post-training method that changes the data the model is trained on and has been shown to be effective in dealing with direct biases. On the other hand, Iterative Nullspace Projection (INLP) is a pre-training method that has shown evidence of being more effective in reducing indirect bias by changing the perception of relationships between words.

In this project, CDA was implemented to mitigate bias in the toxic language classification system. While the interventions proposed and models implemented brought the project closer to achieving its goals, the results indicated that there is still a significant amount of work required to verify some of the results as well as verify the correctness of the implementation. The model was capable of accomplishing its task, but was still in need of debiasing to address the lexical biases towards certain words. The implementation of CDA had a slight impact on the model's performance, but further improvements can be made to the implementation.

Expanding the dataset by not only taking from one source can reduce the lexical biases exhibited by the model. By taking from multiple sources, the biases can be balanced out, reducing the direction in which the task is driven.

Despite the challenges faced, the implementation of a toxic language classification system using BERT was successful and showed promising results. Overall, this study represents a valuable contribution to the field of NLP, highlighting the need for more robust and fair models in detecting and combating hate speech and other forms of toxic language online. There is substantial room for improvement in this field, and this research provides a starting point for further investigations into the realm of debiasing language models with CDA. The experiments should be re-done, and the datasets used should be expanded and inspected more closely to produce more accurate and robust models.

## REFERENCES


[1] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. (2016, July). "*Man is to computer programmer as woman is to homemaker?Debiasing word embeddings*". CoRR, [Online], Available: https://arxiv.org/abs/1607.06520

[2] R. H. Maudslay, H Gonen, R. Cotterell, S. Teufel (Feb, 2020), "*It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution*", CoRR, [Online], Available: https://arxiv.org/pdf/1909.00871.pdf

[3] T. Davidson, D. Warmsley, M. Macy and I. Weber, (2017, March)." *Automated Hate Speech Detection and the Problem of Offensive Language*", CoRR [Online], https://arxiv.org/pdf/1703.04009.pdf

[4] K. Lu , P. Mardziel, F. Wu, P. Amancharia, (2019, May), "Gender Bias in Neural Natural Language Processing", CoRR[Online], Available: https://arxiv.org/pdf/1807.11714.pdf

[5] J. Turian, L. Ratinov, Y. Bengio ( 2010, July), "Word representations: A simple and general method for semi-supervised learning", CoRR [Online], Available: https://aclanthology.org/P10-1040/

[6] J.Lauret, (2019, September), *"Amazon's sexist AI recruiting tool: how did it go so wrong? "* [online], Available: https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

[7] X. Zhou, M. Sap, S. Swayamdipta, N. A. Smith, Y. Choi, (2021, January), "Challenges in Automated Debiasing for Toxic Language Detection", CoRR [Online], Available: https://arxiv.org/pdf/2102.00086.pdf

[8] D. Erhan, Y. Bengio, A. Courville, P. Manzagol, P. Vincent, (2010, May), "Why does unsupervised pre-training help deep learning?.", Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, [Online], Available: https://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf

[9] S. Gonzalez-Carvajal and E. C. Garrido-Merchan, (2020, May), "Comparing BERT against traditional machine learning text classification", CoRR, [Online], Available: https://arxiv.org/abs/2005.13012

[10] A. Caliskan, J. J. Bryson and A. Narayanan (2017, May), "*Semantics derived automatically from language corpora contain human-like biases*", CoRR, [Online], Available: https://arxiv.org/pdf/1608.07187.pdf

[11] K. Lu, P. Mardziel, F. Wu, P. Amancharia. (2019, May), "*Gender Bias in Neural Natural Language Processing*". CoRR, [Online], Available: https://arxiv.org/abs/1807.11714

[12] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, Y. Goldberg.(2020, July). "*Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection*", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, [Online], Available: https://aclanthology.org/2020.acl-main.647/

[13] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, W. Meira Jr, "*Like Sheep Among Wolves: Characterizing Hateful Users on Twitter*", CoRR, [Online], Available: https://arxiv.org/abs/1801.00317

[14] B. van Aken, J. Risch, R. Krestel, A. Löser, (2018, October),"*Challenges for Toxic Comment Classification: An In-Depth Error Analysis*", Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), [Online], Available: https://aclanthology.org/W18-5105/

[15] J. Devlin, M. Chang, K. Lee, K. Toutanova, (2019, May), "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*", Association for Computational Linguistics, [Online], Available: https://aclanthology.org/N19-1423/

[16] C. Sun, X. Qiu, Y. Xu, X. Huang, (2020, February), "How to Fine-Tune Bert For Text Classification", CoRR, [Online], Available: https://arxiv.org/abs/1905.05583

[17] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K. Chang, (2018, June), "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), [Online], Available: https://aclanthology.org/N18-2003/

[18] X. Zhou, M. Sap, S. Swayamdipta, N. A. Smith, Y. Choi, "*Challenges in Automated Debiasing for Toxic Language Detection*", CoRR, [Online], Available: https://arxiv.org/abs/2102.00086