A Scalable Approach to Debiasing Toxic Comment Classifiers using Hierarchical Attention Networks

Arjun Neervannan
arjun.neervannan@gmail.com
University High School
Irvine, California

ABSTRACT

Cyberbullying is a large issue in America, especially through toxic comments. Recent studies demonstrate that machine learning algorithms, particularly text classifiers designed to remove these toxic comments, can exhibit bias towards identity terms. Current bias mitigation techniques use non-transparent models and require manual identification of biased identity terms, which neither scale nor debias sufficiently. The bias detection approach proposed in this paper uses Hierarchical Attention Networks, linguistic rules and grid search to more comprehensively detect bias. The proposed technique is more scalable and comprehensive as it pinpoints where bias exists that can then be filtered automatically and evaluated as opposed to prior approaches that require manual analysis. Our model identified 48% more identity terms that the model was bias against than prior manual approaches. Additionally, on a standard benchmark, we retain accuracy and significantly reduce bias; the overall accuracy remained at 0.98 (AUC) before and after the debiasing processes, but the False Positive Equality Difference (Bias Metric) improved by 44% on the most debiased model. Our bias detection approach can be used to make fairer and more transparent toxic comment detection algorithms and can be applied to other bias mitigation problems.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → *Natural language processing*; • **Applied computing** → *Sociology*.

KEYWORDS

ethics, bias, cyberbullying, toxic language, Natural Language Processing, Hierarchical Attention Network, fairness, transparency, interpretability, explainability

1 INTRODUCTION

Artificial Intelligence (AI) algorithms are increasingly prevalent, and with advanced decision-making capabilities, they are given larger social responsibilities such as picking suspicious activities or individuals in a security footage. However, many do not render fair judgements especially when dealing with socially sensitive topics such as racism, sexism, gender, lifestyle choices and cyberbullying. For example, a study found that AI algorithms display racial bias in image classification problems [4]; another study showed that a hate speech classifier flagged any sentence that contained the word "islam" as toxic [33]. Many recent studies have also scrutinized fairness in machine learning algorithms much more closely [1, 8, 36].

This unintended but harmful bias is often detected much later in the process, after the damage is done. The fairness of the algorithm often goes undiagnosed due to a hidden bias in classification choices or an unbalanced training dataset (i.e. more examples of one type than another) [8] and due to the fact that these models are uninterpretable. Neural net-based AI algorithms learn to model the given data, often work like a black-box, making diagnosing bias a difficult and nuanced task [30].

In this paper we address the decision making capabilities of AI algorithms arbitraging online toxic comments, which is a major catalyst to cyberbullying. It has been widely reported that cyberbullying is a large issue among the youth, with one study reporting that, on average, 7 out of 10 teenagers has been a victim of cyberbullying [18]. Furthermore, another study cited that 22.5% of respondents reported to have been cyberbullied in the prior 30 days through toxic comment or harmful language-based cyberbullying in the form of "mean and hurtful comments online," with an additional 10.1% reporting of cyberbullying in the form of "posting mean comments about my race of color" [13]. Specifically, toxic comments are defined as "rude, disrespectful, or unreasonable comments that are likely to make one leave a discussion" [8]. Cyberbullying has been proven to have many additional negative effects, such as depression and even as extreme as suicide [16].

Many of these toxic comments use identity terms (terms that one may identify by, including but not limited to racial and gender identity terms) in derogatory ways so as to write racist, sexist, or insensitive comments that attack another person [8]. In fact, a study reported that 87% were cyberbullied by race, sexuality, academic achievement, financial status, or religion, which falls under the umbrella of identity terms [7].

AI algorithms are being employed to combat this cyberbullying issue by flagging toxic comments in online forums. However, many AI algorithms designed to filter these comments out often incorrectly associate identity terms with toxicity, removing non-toxic comments that contain these identity terms and suppressing potentially productive discussions on sensitive topics that use identity terms in non-toxic ways [8]. Often censorship is not even attempted and several popular online papers and forums have taken down their online comments section not being able to deal with the vitriolic nature of the comments [27, 28].

Current techniques to mitigate this bias rely largely on manual identification of bias, detection of where bias actually occurs and the use of non-transparent black-box models instead of transparent ones [8]. We believe that this approach is ineffective because detecting bias requires knowledge of where the model is truly biased, which we believe can be done better with a transparent model

(i.e., "listening to the model"). Furthermore, current techniques to mitigate bias can result in reduced model performance [22].

We hypothesized that our method could mitigate and/or eliminate these negative consequences by using a model that provides more transparency, then use a very targeted and surgical approach to detect and stream-line the selection of identity term bias. We hypothesized that using a Hierarchical Attention Network will be more transparent which can then be used to surgically to find the source of bias along with additional linguistic filters. Finally, we fine tune the hyperparameters using grid search and choose optimal set of parameters. This process we believe can at once improve bias detection, scalably debias without negatively impacting the classification accuracy.

To summarize our proposed approach, we developed a scalable bias detection method that used an attention model to pinpoint exactly where bias existed and leveraged linguistic rules to detect identity terms. Additionally, we used a standard debiasing approach of data augmentation to test our approach; we used a grid search method to find the optimal number of terms to augment the dataset.

This project also served as a demonstration of the framework to use a transparent model when debiasing toxic comment classifiers, suggesting that this pipeline could be used in other bias mitigation problems. While this project used automated methods to select the words that the model was biased against, human-based filtering could be introduced into the pipeline in order to make the process more accurate.

2 BIAS IN TEXT CLASSIFICATION

2.1 Bias Definition

Specifically, we focused on detecting unintended bias towards identity terms. We use the definition of "unintended bias" provided in [8] and adapted from [12]; the term "unintended bias" refers to inequity of model performance on different classes or groups of data in a dataset. Unintended bias can be thought of as the "fairness" a model displays across classes or subgroups of data. A truly fair/unbiased model should be equally accurate across all groups of data. In this study, the groups of data referred to comments containing certain identity terms. Therefore, in this experiment, an unbiased model would have equal accuracy across comments containing different identity terms.

In addition, the overall accuracy of the model is not always correlated with the bias. Since bias is a measure of fairness and the model treating all subgroups equally accurately, an inaccurate model may not necessarily be biased, as long as it treats all groups of the data equally inaccurately. However, an accurate model can be very biased if it treats just a few groups of data very poorly but does well on the rest of the dataset. Thus, while we measured the model's accuracy through Area Under Curve (AUC), our primary metric was the Bias Metric (defined in Section 3.5).

The unintended bias noted in prior toxic comment classification models likely occurs because of a dataset imbalance on the identity terms [8]. For example, if a model trained to classify toxic comments was given 100 comments containing the word "gay," but 90% of those were toxic, the model would be likely to attribute the word "gay" with toxicity, even though the word itself is not actually toxic.

2.2 False Positive Bias Definition

False Positive Bias is defined as an unintended bias that occurs when a model flags non-toxic sentences as toxic because they contain certain identity terms [8]. For example, if the model flagged all comments that contained the word "gay" as toxic, regardless of whether the comment was toxic or not, that model would have a false positive bias towards the word "gay."

2.3 Identity Term Definition

An identity term is defined as anything someone may identify by. Many identity terms are used as nouns in certain cases and as adjectives in other cases. As noted above, a common feature of many toxic comments is the use of identity terms in derogatory ways. An example of this linguistic feature is shown below in Table 1. In the first sentence, the word "muslim" is used as an adjective, while in the second sentence, the word is used as a noun. This linguistic feature of identity terms can be used generate a proxy list.

2.4 Bias Mitigation

While bias mitigation was not the primary goal of this paper, we tested our novel bias detection approach by running a standard data augmentation bias reduction technique.

A common bias mitigation strategy used is data augmentation [6, 8]. Specifically, this technique involves adding more examples that contain a particular term to the training dataset, and retraining the model on this augmented dataset. By providing more data, biases can be broken as the model is given examples with the particular identity term in different contexts.

Since we targeted false positive bias specifically, we chose to augment the dataset with only non-toxic examples. False positive bias occurs when the model flags a comment as toxic because of a specific term; therefore, we hypothesized that adding non-toxic comments would give the model more examples of identity terms used in non-toxic contexts, and would thus mitigate its false positive bias.

We added non-toxic examples that contained the identity terms that the model was biased against. We used data from the Wikipedia Corpus of Text to generate these non-toxic examples. [8] confirmed that 99.5% of the sentences in this dataset were non-toxic by examining a 1000-sentence sample; therefore, we were able to use this corpus of text to generate non-toxic examples.

2.5 Model Transparency

In order to make bias detection and mitigation more targeted, we used a transparent hierarchical attention model to specifically point out the identity terms that the model was biased against [32, 35]. By having a transparent model, we were able to better understand the model's properties (through the use of attention weights, explained in Section 3.1.3). Recently, regulatory government agencies have mandated that entities explain their predictions [10]. Recent work has been done to create more explainable NLP models [20, 21, 23].

2.6 Project Focus

Specifically, this project focused on detecting false positive bias in a more scalable way. We focused on detecting only identity term

Sentence	Identity Term	Part of Speech
`There were many Muslims present at the mosque on Sunday.`	Muslim	Noun
`There were many Muslim scientists in the 10th century.`	Muslim	Adjective

Table 1: Table showing how Identity Terms can often be used as noun or adjective depending on the context in the sentence - Linguistic Feature

bias and not other sources of bias. However, given the nature of the framework, we believe that this method could be applied to detecting other biases with toxic comment classification or even other text classification problems.

3 METHODOLOGY

The pipeline has five main steps: baseline training, identity term generation, bias detection, grid search debiasing (bias mitigation), and control model training. The first three steps were used to detect bias, while the last two steps were used to show bias mitigation on the approach.

3.1 Baseline Training

This paper used a Hierarchical Attention Network to pinpoint the terms that the model was biased against [35]. The intuition behind the Hierarchical Attention model is that words form sentences and sentences form the comments and the importance of the words in a comment are context dependent, i.e. the same word or sentence may be differentially important in different contexts [35]. This provides the basis for the context sensitive attention weights thus providing transparency into the model further aiding the debiasing process.

3.1.1 The Model. This experiment used a Hierarchical Attention Network that had a word encoder composed of an input layer, and embedding layer which used the 840-billion 300-dimensional GloVe vectors [25], a Bidirectional LSTM layer with 100 nodes [14], a Dense layer with 200 nodes, and the Attention Layer. This was followed by a sentence decoder, which was composed of a Bidirectional LSTM [14] with 100 nodes, a Dense Layer with 200 nodes, an Attention Layer, and a Dense Layer with Softmax activation to output the binary classification for each example. The recurrent layers made the model insensitive to input comment length (shorter comments did not have to be padded).

3.1.2 The Dataset. The training dataset, which was used from the Wikipedia Toxic Comment Dataset (as in [8]), contained over 99,000 labeled comments (toxic/non-toxic), about 90% of which were non-toxic and the remaining toxic (an imbalanced dataset). The dataset was annotated by human raters, where toxicity was defined as a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" [8]. The dataset was preprocessed and tokenized into 3D arrays [34]. The validation and test datasets had over 30,000 comments each, both of which contained a similar 90/10 split on non-toxic/toxic data. The baseline model was then trained for 5 epochs on the training data.

3.1.3 Attention Weights. To measure the importance of a word, we used a metric called the attention weight, which we derive from the Hierarchical Attention Network [35]. The attention weight specifically measures the "informativeness" of the word in the context [35]. Using this attention weight, we were able to detect which identity terms the model considered to be important in false positive classifications, allowing us to later on mitigate this bias in a very targeted way.

An example of a sentence with the accompanying attention weights is shown below in Table 2. The sentence is toxic, and the words “dumb” and “idiot” have the highest attention weights as they are the highest contributors to toxicity in the sentence.

3.2 Identity Term Generation

To generate a proxy list of identity terms, a list of terms in the English language that are used as nouns and adjectives in different contexts is generated (using the linguistic feature described in Table 1). This allows us to have a starting point of identity terms to later filter down.

In this experiment, additional filters were run to narrow down the initial list of identity terms and remove noisy terms. These filters included:

- Inverse Document Frequency (IDF) - the IDF measures the frequency of words; words with a lower IDF score were used less frequently. This was done to eliminate common words. The calculation for IDF is shown below. N is the total number of comments, and df is the number of comments that contain one usage of the term.
- Frequency of word count - This step was done to eliminate words that did not have enough occurrences in the validation dataset.
- English language dictionary checker - Some words picked up by the identity term generation process were not English words, such as usernames or other random words. Removing these words reduced the amount of noise.

In future applications, a human reviewer can be introduced to manually narrow down the list of identity terms. This is more efficient than current methods as the human reviewer would only have to process a shortened list of around 300-400 words to eliminate unimportant words instead of manually generating the terms.

3.3 Bias Detection

Additionally, the mean attention weights for false positive classifications were assigned to each identity term. This made clear which terms the model was specifically biased against, allowing for the framework to pinpoint the debiasing process; higher attention weight for an identity term indicated that the model was more biased towards that term. The use of attention weights also allows human-based filtering to more efficiently and quantitatively

Description	Word 1	Word 2	Word 3	Word 4	Word 5	Total
Word	You	are	a	dumb	idiot.	-
Attention Weight	0.01	0.02	0.01	0.45	0.51	**1.00**

Table 2: Shows the attention weight assignment by the Hierarchical Attention Model for a given sentence. Sum of the attention weights for the words in a sentence is equal to 1. We particularly use the attention weights from the toxic sentences for the false positive scenarios for our debiasing process.

uncover which words the model may be biased against. This key improvement from other papers allows for a more accurate and scalable way of debiasing since the model's weights are used to directly diagnose bias.

3.4 Grid Search Debiasing

The debiasing procedure involved augmenting data with "n" non-toxic comments containing each identity term to help the model improve its understanding of context of identity terms. Augmenting data allows the model to break relationships between identity terms and toxicity, and the model becomes more unbiased towards a certain term. This technique has been used before by other researchers. We used non-toxic comment data from the Wikipedia Corpus of Text to augment the original dataset. As part of the grid search process we vary the number of sentences that we add for each identity term (hyperparameter of the model).

Using the identity term list generated in Section 3.2, a grid search iterating over the number of comments augmented per term and the percent of identity terms selected from the original list (hyperparameter of the model) was run. The model with the highest improvement in False Positive Equality Difference (FPED, explained in section 3.6) was selected to be the most optimal model. Searching over the entire hyperparameter space ensured that the most optimal model was selected. Additionally, to minimize training variance, the entire grid search process was repeated a total of 5 times and the results were averaged across trials.

The grid search was iterated over the following hyperparameter options:

- **Percent of Identity Terms (ranked by mean attention weight):** 90%, 80%, 70%, 60%, 50%, 40%.
- **Number of non-toxic comments added per identity term:** 10 comments, 20 comments, 30 comments, 40 comments

3.5 Control Model

To ensure that the model's bias was not improving solely because of data augmentation, an additional control model for the most optimal debiased model was trained by augmenting data with non-toxic comments that did not contain identity terms. Adding random comments to the dataset and retraining the model would confirm if the bias improvement occurred because of the mere addition of data or if it occurred because of word selection. All of the comments added were confirmed to not have any identity terms. The bias improvement on this model was compared with the most optimal model.

3.6 Evaluation Metrics

To measure the performance of the model, 4 metrics were used:

- Mean Attention Weights - These weights are generated from Hierarchical Attention Network and indicate the importance of the word in determining the classification. This was used to determine the words that model was biased against. A higher attention weight indicates that the model paid more attention to that word when classifying the sentence incorrectly, which points to a potential source of bias. Since we targeted false positive bias in identity terms, only the weights from comments with a false positive classification were used. The total sum of all the attention weights of the words in a sentence is 1; therefore, the maximum attention weight of a word is theoretically 1, and the minimum is 0.
- False Positive Rate Equality Difference (FPED) - This metric was derived from the Equality of Odds concept presented in [12], which is satisfied when the False Positive Rate and the False Negative Rate are equal. Furthermore, this metric was used in prior research papers, including [8] and [24].
  FPED measures the equality of the False Positive Rate across different identity terms. This follows from the definition of unintended bias; model should be equally "accurate" across various terms. We did not measure the False Negative Rate Equality Difference because we were focused on specifically mitigating False Positive Bias.
  The False Positive Equality Difference (FPED) is calculated as shown below. FP_t represents the number of false positives for identity term t, TN_t represents the number of true negatives for identity term t, FPR represents the overall false positive rate, and FPR_t represents the False Positive rate for identity term t.

$$FPED = \frac{1}{t}\sum |FPR - FPR_t| \tag{1}$$

- FPED improvement ($\text{Bias}_{\%Change}$) - This metric measures the percent improvement in FPED between the baseline model and the debiased models from the grid search. The model with the highest $\text{Bias}_{\%Change}$ was the most debiased model.
  The calculation for $\text{Bias}_{\%Change}$ is shown below. $\text{FPED}_{baseline}$ represents the FPED for the baseline model, and $\text{FPED}_{debiased}$ represents the FPED for the model that is being debiased.

$$Bias_{\%Change} = \frac{FPED_{debiased} - FPED_{baseline}}{FPED_{baseline}} \times 100 \tag{2}$$

- AUC on the ROC - This metric was used to measure overall accuracy. Specifically, this metric was used to ensure that there was not a drop in accuracy after the debiasing process.

4 EXPERIMENTAL DESIGN

The experimental design includes training the baseline model, performing linguistic filtering (with the noun/adjective filter) to pick the identity terms, grid searching to determine the optimal set of hyperparameters - percent of identity terms to use and number of sentences per identity term to augment the dataset with for the debiased model. The control model is trained with these same set of hyperparameter values to show that the debiasing is not the effect of data augmentation.

5 RESULTS

Table 3 shows a small selection of the identity terms picked by the baseline model. In addition, the model identified many non-toxic identity terms that were used in toxic ways in comments, such as “homosexual”, “slavs”, “semite”, and others. In this case, these words were "highly paid attention to" in false positive classifications, suggesting that the model was biased towards those words. Using the noun-adjective filter as well as the attention weights, the model was able to self-diagnose its bias and accordingly add neutral comments containing these terms to debias.

Identity term	Mean FP Attn Weight from Baseline
canadian	0.423
semite	0.407
inbred	0.372
homosexual	0.323
slavs	0.312
maoist	0.303

Table 3: Shows a small selection of the identity terms along with their mean attention weights for the baseline model on the validation dataset. These weights are generated before the debiasing process. Mean False Positive Attention Weights are computed by averaging attention weights for multiple occurrences of the term in the validation dataset for false positives.

While this list of words was generated automatically by selecting noun-adjective identity term words and filtering through Inverse Document Frequency (IDF), frequency, and English language filters, a future application could instead use a human to remove insignificant or noisy terms.

In total, 358 identity terms were generated from the model, and the debiasing procedure was run on all terms in this list. However, a manual evaluation revealed that 74 of these terms were actually identity terms and the remaining 284 were noisy terms. A prior paper, which instead manually generated a list of identity terms, only included 50 identity terms [8]. Our paper thus improved on the prior paper by generating 48% more identity terms that the model was biased against.

$Bias_{\%Change}$ from Grid Search on Validation Dataset for Model Selection					
% of Identity Terms Selected	Number of Comments Augmented per Term in Validation Dataset				
	Baseline	10 sentences	20 sentences	30 sentences	40 sentences
90%	0.0%	-15.7%	11.6%	32.1%	12.8%
80%	0.0%	5.0%	32.6%	0.6%	17.2%
70%	0.0%	9.2%	18.3%	21.2%	3.6%
60%	0.0%	28.8%	44.1%	30.6%	12.9%
50%	0.0%	12.1%	13.9%	-2.6%	15.8%
40%	0.0%	26.9%	13.9%	32.7%	22.8%

Figure 1: Grid Search Results - Shows the percent improvement in the FPED for the different debiased models over the baseline model (rows - percent of identity terms selected, columns - number of comments added per term) averaged over 5 trials on validation dataset. The model with 60% of terms selected and 20 sentences added per term (in dark blue) is the most effective as it debiased the most.

Table 4 shows the percent improvement in the FPED for the different debiased models over the baseline model (rows - percent of identity terms selected, columns - number of comments added per term) averaged over 5 trials on validation dataset. The model with 60% of terms selected and 20 sentences added per term (in dark blue) is the most effective as it debiased the most.

Model	AUC on Test Dataset	$Bias_{\%Change}$
Baseline	0.982	-
Debiased-60/20	0.982	44.1%
Control-60/20	0.983	0%

Table 4: Table showing comparison of the Baseline Model, Debiased Model with 60% of identity terms and 20 sentences augmented per identity term and the Control Model with the same 60/20 combination. Accuracy did not change for the models but there is a significant FPED improvement proving that we were able to successfully debias. The numbers for the Control Model show that the improvement seen in the Debiased Model was not the effect of data augmentation.

Table 4 shows the AUC computed on Baseline model, AUC on Debiased Model with 60% of Identity Terms and 20 sentences added per term and AUC of the corresponding 60/20 Control Model.

The optimal debiased model (60% of identity terms and 20 comments per term) attained an average $Bias_{\%Change}$ of 44 %, as compared to the control model, which did not improve at all (0% $Bias_{\%Change}$). This indicated that the improvement seen in the debiased model was not a result of data augmentation. In addition, the overall accuracy (AUC) of the debiased model was almost exactly the same as the baseline model, proving that the debiasing process was non-invasive and targeted as the accuracy did not reduce at all. These two data points demonstrate the efficacy of the approach and its ability to strategically and carefully remove biases without disrupting the model. This is an important improvement as this ensures a fair and and improved classification of toxic comments while removing the bias. This is the core improvement of

this approach as this directly affects the ability to use identity terms in non-toxic ways.

As shown in Figure 3, the model achieved a significant FPED Improvement from the baseline on many identity terms, including "young", "christian", "ethnic", "german", and others. This again showed the efficacy of the approach to detect potential identity term biases by itself and surgically fix those biases. This graph shows evidence that the model is able to target specific identity terms that it is biased against and mitigate them.

Figure 4 shows examples from the dataset that highlight the model's ability to detect and mitigate its biases. The intensity of the highlights indicate the extent to which the model "paid attention" to the word (context sensitive attention weight).

In the first sentence, before debiasing, the model predicted the sentence to be toxic, even though it was non-toxic. The model paid attention to the identity term "racist" in this sentence and considers it to be an indicator of toxicity. After the debiasing process, however, the model understood the context of the sentence better and classified the sentence correctly as non-toxic. While the model still pays attention to the identity term "racist," it no longer considers it to be a toxic term as the sentence itself is classified as non-toxic; the attention paid to a word to in a non-toxic sentence is an indicator of that word's effect on its "non-toxicity" as opposed to the toxicity.

In the second sentence, before debiasing, the model predicted the sentence to be toxic, even though it was non-toxic, and paid some attention to the identity term "gay." After debiasing, the model correctly predicted the sentence to be non-toxic, and even pays more attention to the identity term "gay." The model can read the context better after debiasing and is able to understand that the word "gay" can be used in a positive light as well.

In the third sentence, before debiasing, the model predicted the sentence to be toxic, which was correct. However, the attention weight was placed primarily on "muslim," as opposed to "ass." In this case, the word "ass" was a key trigger for the toxicity, and not the word "muslim." After debiasing, the model still classified it correctly, but shifted the attention weight primarily to the word "ass." While there was still some attention being paid to the word "muslim," was is significantly better than before debiasing.

These three examples showed the efficacy of the framework in detecting and mitigating identity term bias in a real-life setting. These sentence examples are corroborated with evidence regarding the overall reduction in bias, proving that this framework was effective in detecting and mitigating bias.

6 RELATED WORK

As per the study conducted by the Urban Institute [38], rates of cyberbullying among youth may be increasing as access to technological devices increases, and may be highest among the most vulnerable youth populations (e.g., females, LGBTQ). A study from UK-based nonprofit Ditch the Label in 2013 reported that 7 out of 10 teenagers was a victim of cyberbullying [18]. A study from Swansea University also reported that victims of cyberbullying were twice as likely to attempt suicide and self-harm [16].

Another study reported that 22.5% of respondents were cyberbullied in the prior 30 days through "mean or hurtful comments online," 12.7% through others "posting mean names or comments online with a sexual meaning," 12.2% through others "threatening to hurt them online," 11.9% through others "threatening to hurt them through a text message," along with many others, which we classified as forms of toxic comment-based cyberbullying [13]. Additionally, a study reported that 87% were cyberbullied by race, sexuality, academic achievement, financial status, or religion, which we classified as forms of identity term-based cyberbullying [7].

[3] addressed gender stereotyping issues by debiasing word vectors but did not address all other types of unintended bias that we now see in machine classifications. [37] also tried to address gender bias by calibrating existing prediction models for multilabel classifications but did not provide a debiasing methodology.

[5] and [11] explored fair machine learning by pre and post-processing training datasets essentially dealing with measuring algorithmic bias but did not provide ways of debiasing.

Recent work has also been done regarding auditing black-box models and making NLP models more transparent. [21] focused on making a widely applicable method to introduce *compositionality*, or "building sentence meaning from the meanings of words and phrases" [21]. [20] proposed introducing *rationales*, or pieces of input text that serve as a justification for a prediction, into NLP models. [23] introduced *contextual decomposition*, a method for analyzing the individual predictions made by LSTM models [14], without actually changing the model. [35] proposed the Hierarchical Attention Model, which would use "attention weights" to explain the predictions of the model. This was the primary model used in this paper. Furthermore, a study showed evidence that regulatory agencies require entities to explain their predictions [10].

Studies have shown the possible ways that bias can occur as well as possible ways to measure this bias [9, 12]. [12] proposed the Equality of Odds definition of bias, which we used to derive the False Positive Equality Difference (FPED) definition. [9] proposed that bias mainly occurred through three main ways: preexisting bias (bias that exists with or without the model), technical bias (bias that exists because of limitations in the computing power), and emergent bias (bias that results after the model has been created from its interactions with real-world data). Our study is different in that we do not propose a new way to measure bias but rather show the efficacy of our novel method of debiasing.

Studies have shown that bias can occur in non-text settings and have proposed ways to mitigate it [2, 4, 15, 29, 31]. This study focuses on bias in text-based classifiers rather than other contexts and proposes novel automatic methods to mitigate bias.

Prior methods have shown that gender bias can exist in text-based algorithms and have used manual methods to mitigate this bias [3, 26]. [3] specifically adjusted the word embeddings to fix the gender bias. [26] adjusted the loss function in order to mitigate gender bias in NLP models.

[24] also tried to address gender bias for abusive language detection models by debiasing word vectors, augmenting more data and changing model architecture. Their results seem to show promise for removing gender bias but their method does not scale for other identity dimensions such as race and religion.

Our project does not focus on gender bias specifically, but rather identity term bias in toxic comment classifiers; furthermore, we

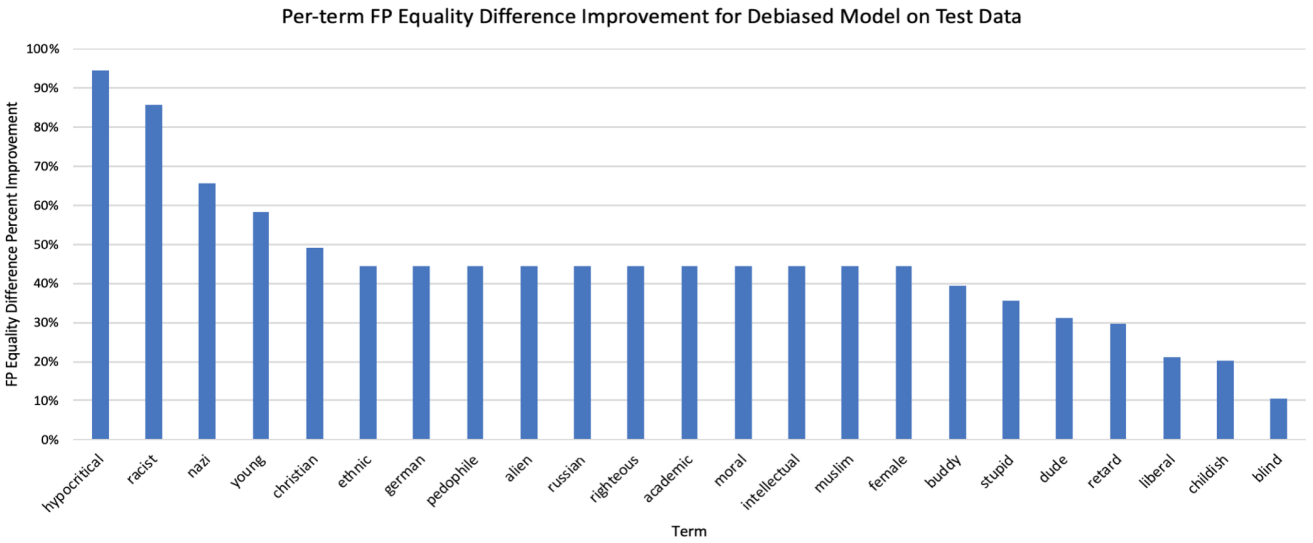

Figure 2: Graph showing the per-term False Positive Equality Difference (FPED) improvement for the identity terms between the optimal debiased model and the baseline on Test Data.

Baseline [actual: non-toxic; pred: toxic]
Disagreeing with you and your political views does not make someone a racist.

Debiased-60/20 [actual: non-toxic ; pred: non-toxic]
Disagreeing with you and your political views does not make someone a racist.

Baseline [actual: non-toxic; pred: toxic]
In the UK, as of March 2006, 80% of aids sufferers have contracted it from sex between gay males.

Debiased-60/20 [actual: non-toxic ; pred: non-toxic]
In the UK, as of March 2006, 80% of aids sufferers have contracted it from sex between gay males.

Baseline [actual: toxic ; pred: toxic]
Get your muslim ass out of here.

Debiased-60/20 [actual: toxic ; pred: toxic]
Get your muslim ass out of here.

Figure 3: Sentence examples from the Wikipedia Comment Dataset with highlighting intensity showing the impact that the word had on the classification; thus proving the model to be interpretable. In some of these scenarios the baseline model incorrectly outputted a comment as toxic but the debiased model outputted it correctly. The highlights indicate the extent to which the model "paid attention" to the word (context sensitive attention weight). In all of the examples, the model was not only able to classify the sentence correctly but was also assign appropriate attention weights.

use an automated method using an attention network instead of a black-box model.

A study released in 2018 by [8] at Google AI showed the bias that toxic comment detection algorithms displayed towards a list of 51 hand-selected identity terms by using a metric, the pinned Area Under Curve (pAUC), to compare the performance of the model on the subset of the comments that contained each of those identity terms to that on the dataset as a whole. The study also debiased the model by adding more non-toxic data that contained the identity terms to the training dataset. Despite the remarkable findings and novel procedure, the study displayed some fundamental flaws that this paper addressed. The first and most important flaw was the unsupported assumption that the 51 hand-selected terms were words that the model was biased against. These words were selected mostly based off of human intuition and the fact that the terms were "identity terms." Rather than relying on pure intuition, this paper uses a quantifiable metric that clearly shows whether or not the model is actually biased towards a certain term. The second flaw in the model was its lack of interpretability. The authors of the prior paper made the unsupported assumption that the model considered their list of 51 hand-selected terms to be important, and without any proof to show otherwise, it was possible that the model was not actually biased against these terms to begin with. This project addresses that issue by introducing interpretability to the model by using the attention-based sequence learning model, which explains which words the algorithm considered important when classifying a sentence a certain way. Another flaw in the prior model is that the Convolutional Neural Networks (CNN) that it used were sensitive to the length of the comments (toxic comments tend to be shorter) [17, 19]. As a result, the comment length had to be balanced, requiring additional human intervention [8].

Another study that expanded [8] adjusted the loss function and used feature attribution to allow for transparency [22]. This study also required manual methods to fix the bias, which is less scalable than our method. Additionally, [8, 22] also used a Convolutional Neural Network for the classification [17], which was sensitive to comment length, unlike the Hierarchical Attention Networks used in our paper, which used Long-Short Term Memory Networks (LSTMs) [14, 35]. The LSTM networks were less sensitive to comment length and were thus more scalable [14].

Current debiasing approaches to mitigate bias in identity terms depend on a manual approach in detecting biased identity terms

and are often uninterpretable models [8]. Our paper is different in that bias detection is done automatically by using transparent models, rather than using manual methods to identify bias.

7 DISCUSSION

The results of this study show that the debiasing process worked as the model was able to significantly reduce bias. We analyze the results in the context of the whole pipeline below and compare each step to prior work.

7.1 Bias Detection

The first step in the pipeline was to detect the words that the model was biased against by first generating a list of identity terms through the noun-adjective criteria, then assigning the attention weights to each of these terms (sourced from the trained Hierarchical Attention Model), which would signify how biased the model was towards that identity term. In this experiment, automatic filters were used to remove noisy terms, but in future applications, a human could be used to filter down these terms.

The bias detection process was fairly accurate. As shown in Table 3, the process was able identify many identity terms that the model was biased against, and show the "intensity" of the bias through the attention weight.

This improvement was a key one over prior papers. By displaying the severity of the bias for each word, our bias detection is far more accurate, especially over methods that used human intuition and other manual methods to generate identity terms that the model was biased against [8]. Even with human intervention, we believe that our method would still be more accurate and efficient than prior methods; rather than generating terms from scratch, our method would only require the human reviewer to filter out noisy words, and would look at the various metrics, such as the validity of the identity term (whether or not it is an identity term) and the attention weight, to determine whether or not to eliminate the word quantitatively rather using qualitative terms. This is a far more scalable and repeatable process than generating words.

7.2 Grid Search Debiasing

The second step in the pipeline was to search over the hyperparameter space to determine the optimal number of identity terms to debias and the optimal number of non-toxic sentences to add per term. As explained in Section 3, the non-toxic sentences were sourced from the Wikipedia Corpus of Text.

As shown in Figure 1, the optimal combination of hyperparameters was 60% of identity terms selected (with the identity term list ranked by false positive mean attention weight) and 20 comments added per term. Prior methods did not use grid search to determine this optimal amount of data to add [8]. Using grid search for data augmentation allows our model to scale better to other contexts because it automatically determines the optimal hyperparameter combination, as opposed to manually determining the optimal hyperparameters.

7.3 Control Model

The third step in the pipeline was to train the control model by augmenting the same number of random non-toxic comments as the top debiasing model. This would confirm if the bias improvement was a result of the model having more data in general to train on, or if it was a result of the model having more data for the selected identity terms.

As shown in Table 4, the $\text{Bias}_{\%Change}$ was 0%, as opposed to the 44% figure for the most debiased model. This was evidence that the bias improvement was not a result of mere data augmentation, but rather a targeted improvement by the pipeline that reduced bias on the specific identity terms. Additional evidence of this targeted bias reduction is shown in Figure 2. The model was able to reduce the FPED for those selected identity terms without losing any accuracy (Table 4).

Our method of control model evaluation was not different than that of [8]. However, our results show that the model was able to target the bias and mitigate those biases in a targeted and surgical manner without disrupting the overall accuracy, which can sometimes be an issue in bias mitigation [22].

8 CONCLUSION AND FUTURE WORK

The project succeeded in its goal, which was to create an advanced, more scalable and an automated debiasing methodology, thus creating a fairer and more accurate and just AI model that allows moderators to facilitate productive online discussions without over censoring. The use of attention networks, unique linguistic based filtering methods, and grid search made the model reliable, scalable and more automated. The project also addressed the comment length limitation, hand selection of identity terms and other limitations in [8]. The debiasing process can also be used in conjunction with manual filtering steps to improve the quality of bias detection and mitigation.

While this approach displayed much promise in automatically detecting bias and debiasing AI models, an area of improvement was the choice of attention words, as the model occasionally generated unimportant words despite the filters. There is also more scope for including, additional types of words to be debiased (instead of just noun-adjective). The model has much promise to be used in multiple languages and in real-world applications, ensuring a fairer and more just platform for online discussions.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. [n. d.]. Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks. Retrieved May 23, 2016 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *CoRR* abs/1707.00061 (2017). arXiv:1707.00061 http://arxiv.org/abs/1707.00061

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *CoRR* abs/1607.06520 (2016). arXiv:1607.06520 http://arxiv.org/abs/1607.06520

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. TeX Users Group, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[5] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[6] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? *CoRR* 1805.12002v2 (2018). arXiv:1805.12002 https://arxiv.org/abs/1805.12002v2

[7] The Futures Company. [n. d.]. 2014 Teen Internet Safety Survey. Retrieved 2014 from https://www.cox.com/content/dam/cox/aboutus/documents/tween-internet-safety-survey.pdf

[8] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification.

[9] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[10] Bryce Goodman and Seth Flaxman. 2065. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, Vol 38, No 3, 2017* (2065). arXiv:1506.01066 https://arxiv.org/abs/1606.08813v3

[11] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 2125–2126. https://doi.org/10.1145/2939672.2945386

[12] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *CoRR* abs/1610.02413 (2016). arXiv:1610.02413 http://arxiv.org/abs/1610.02413

[13] Sameer Hinduja and Justin W. Patchin. 2008. *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying*. Corwin Press, Incorporated.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[15] Heinrich Jiang and Ofir Nachum. 2019. Identifying and Correcting Label Bias in Machine Learning. *CoRR* abs/1901.04966 (2019). arXiv:1901.04966 http://arxiv.org/abs/1901.04966

[16] Ann John, Alexander Charles Glendenning, Amanda Marchan, Paul Montgomery, Anne Stewart, Sophie Wood, Keith Lloyd, and Keith Hawton. 2018. Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review. *J Med Internet Res* 20, 4 (19 Apr 2018), e129. https://doi.org/10.2196/jmir.9044

[17] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[18] Ditch The Label. [n. d.]. The Annual Cyberbullying Survey. Retrieved 2013 from https://www.ditchthelabel.org/wp-content/uploads/2016/07/cyberbullying2013.pdf

[19] Yann Lecun, LÃĺon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*. 2278–2324.

[20] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 107–117. https://doi.org/10.18653/v1/D16-1011

[21] Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2015. Visualizing and Understanding Neural Models in NLP. *CoRR* abs/1506.01066 (2015). arXiv:1506.01066 http://arxiv.org/abs/1506.01066

[22] Frederick Liu and Besim Avci. 2019. Incorporating Priors with Feature Attribution on Text Classification. *CoRR* abs/1906.08286 (2019). arXiv:1906.08286 http://arxiv.org/abs/1906.08286

[23] W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. *CoRR* abs/1801.05453 (2018). arXiv:1801.05453 http://arxiv.org/abs/1801.05453

[24] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. *CoRR* abs/1808.07231 (2018). arXiv:1808.07231 http://arxiv.org/abs/1808.07231

[25] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

[26] Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. *CoRR* abs/1905.12801 (2019). arXiv:1905.12801 http://arxiv.org/abs/1905.12801

[27] Popular Science. [n. d.]. Why We're Shutting Off Our Comments. Retrieved September 24,2013 from https://www.popsci.com/science/article/2013-09/why-were-shutting-our-comments/

[28] Pacific Standard. [n. d.]. JUST KILL ALL OF THE COMMENTS ALREADY. Retrieved August 12, 2014 from https://psmag.com/environment/just-kill-comments-already-88188

[29] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *CoRR* abs/1301.6822 (2013). arXiv:1301.6822 http://arxiv.org/abs/1301.6822

[30] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, NY, USA, 303–310. https://doi.org/10.1145/3278721.3278725

[31] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 53–59. https://doi.org/10.18653/v1/W17-1606

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[33] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[34] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex Machina: Personal Attacks Seen at Scale. *CoRR* abs/1610.08914 (2016). arXiv:1610.08914 http://arxiv.org/abs/1610.08914

[35] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 1480–1489. https://doi.org/10.18653/v1/N16-1174

[36] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *CoRR* abs/1801.07593 (2018). arXiv:1801.07593 http://arxiv.org/abs/1801.07593

[37] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *CoRR* abs/1707.09457 (2017). arXiv:1707.09457 http://arxiv.org/abs/1707.09457

[38] Janine M. Zweig, Meredith Dank, Jennifer Yahner, and Pamela Lachman. 2013. The rate of cyber dating abuse among teens and how it relates to other forms of teen dating violence. *Journal of Youth and Adolescence* 2013 Jul (2013). arXiv:1801.07593 https://www.ncbi.nlm.nih.gov/pubmed/23412689