Detecting Unintended Social Bias in Toxic Language Datasets

Nihar Sahoo[*], **Himanshu Gupta**[*], **Pushpak Bhattacharyya**
CFILT, Indian Institute of Technology Bombay, India
{nihar, himanshug, pb @cse.iitb.ac.in}

Abstract

Warning: This paper has contents which may be offensive, or upsetting however this cannot be avoided owing to the nature of the work.

With the rise of online hate speech, automatic detection of Hate Speech, Offensive texts as a natural language processing task is getting popular. However, very little research has been done to detect unintended social bias from these toxic language datasets. This paper introduces a new dataset *ToxicBias* curated from the existing dataset of Kaggle competition named "Jigsaw Unintended Bias in Toxicity Classification". We aim to detect social biases, their categories, and targeted groups. The dataset contains instances annotated for five different bias categories, viz., *gender, race/ethnicity, religion, political, and LGBTQ*. We train transformer-based models using our curated datasets and report baseline performance for bias identification, target generation, and bias implications. Model biases and their mitigation are also discussed in detail. Our study motivates a systematic extraction of social bias data from toxic language datasets. All the codes and dataset used for experiments in this work are publicly available[1].

1 Introduction

In the age of social media and communications, it is simpler than ever to openly express one's opinions on a wide range of issues. This openness results in a flood of useful information that can assist people in being more productive and making better decisions. According to statista [2], the global number of active social media users has just surpassed four billion, accounting for more than half of the world's population. The user base is expected to grow steadily over the next five years. Various studies (Plaisime

---

[*]These authors contributed equally to this work
[1]https://github.com/sahoonihar/ToxicBias_CoNLL_2022
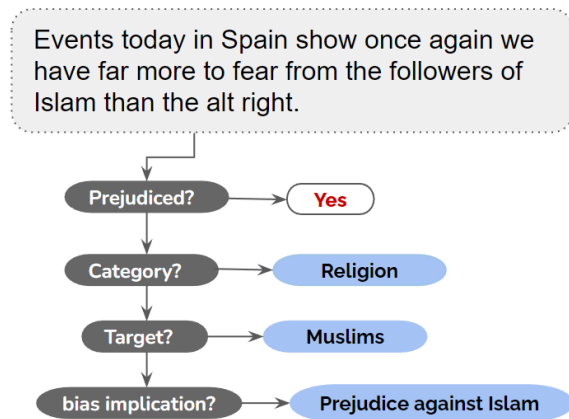[2]https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Figure 1: An illustrative example of *ToxicBias*. During the annotation process, hate speech/offensive text is provided without context. Annotators are asked to mark it as biased/neutral and to provide category, target, and implication if it has biases.

et al., 2020) say that children and teenagers, who are susceptible, make up a big share of social media users. Unfortunately, this increasing number of social media users also leads to an increase in toxicity (Matamoros-Fernández and Farkas, 2021). Sometimes this toxicity gives birth to violence and hate crimes. It does not just harm an individual; most of the time, the entire community suffers as due to its intensity.

We have different perspectives based on race, gender, religion, sexual orientation, and many other factors. These perspectives sometimes lead to biases that influence how we see the world, even if we are unaware of them. Biases like this can lead us to make decisions that are neither intelligent nor just. Furthermore, when these biases are expressed as hate speech and offensive texts, it becomes painful for specific communities. While some of these biases are implied, most explicit biases can be found in the form of hate speech and offensive texts.

The use of hate speech incites violence and sometimes leads to societal and political instability.

BLM (Black Lives Matter) movement is the consequence of one such bias in America. So, to address these biases, we must first identify them. While the concepts of Social Bias and Hate Speech may appear to be the same, there are subtle differences.

This paper expands on the above ideas and proposes a new dataset *ToxicBias* for detecting social bias from toxic language datasets. The main contributions can be summarized as follows:

- To the best of our knowledge, this is the first study to extract social biases from toxic language datasets in English.

- We release a curated dataset of 5409 instances for detection of social bias, its categories, targets and bias reasoning.

- We present methods to reduce lexical overfitting using counter-narrative data augmentation.

In the following section we discuss various established works which are aligned with our work. Section 3 provides information about our dataset, terminology, annotation procedure, and challenges. In section 3, we describe our tests and results, followed by a discussion of lexical overfitting reduction via data augmentation in section 5. Section 6 discusses the conclusion and future works.

2 Related Work

Offensive Text: Unfortunately, offensive content poses some unique challenges to researchers and practitioners. First and foremost, determining what constitutes abuse/offensive behaviour is difficult. Unlike other types of malicious activity, e.g., spam or malware, the accounts carrying out this type of behavior are usually controlled by humans, not bots (Founta et al., 2018).The term "offensive language" refers to a broad range of content, including hate speech, vulgarity, threats, cyberbully, and other ethnic and racial insults (Kaur et al., 2021). There is no single definition of abuse, and phrases like "harassment," "abusive language," and "damaging speech" are frequently used interchangeably.

Hate Speech: Hate Speech is defined as speech that targets disadvantaged social groups in a way that may be damaging to them. (Davidson et al., 2017). Fortuna and Nunes (2018) defines Hate speech as follows: "Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used".

Bias in Embedding: The initial works to explore bias in language representations aimed at detecting gender, race, religion biases in word representations (Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019). Some of recent works have focused on bias detection from sentence representations (May et al., 2019; Kurita et al., 2019) using BERT embedding.

In addition, there have been a lot of notable efforts towards detection of data bias in hate speech and offensive languages (Waseem and Hovy, 2016; Davidson et al., 2019; Sap et al., 2019; Mozafari et al., 2020). Borkan et al. (2019) has discussed the presence of unintended bias in hate speech detection models for identity terms like islam, lesbian, bisexual, etc. The biased association of different marginalized groups is still a major challenge in the models trained for toxic language detection (Kim et al., 2020; Xia et al., 2020). This is mainly due to the bias in annotated data which creates the wrong associations of many lexical features with specific labels (Dixon et al., 2018). Lack of social context of the post creator also affect the annotation process leading to bias against certain communities in the dataset (Sap et al., 2019).

Social bias datasets: More recently, many datasets (Nadeem et al., 2021; Nangia et al., 2020) have been created to measure and detect social biases like gender, race, profession, religion, age, etc. However, Blodgett et al. (2021) has reported that many of these datasets lack clear definitions and have ambiguities and inconsistencies in annotations. A similar study have been done in (Sap et al., 2020), where dataset has both categorical and free-text annotation and generation framework as core model.

There have been few studies on data augmentation (Nozza et al., 2019; Bartl et al., 2020) to decrease the incorrect association of lexical characteristics in these datasets. Hartvigsen et al. (2022) proposed a prompt based framework to generate large dataset of toxic and neutral statements to reduce the spurious correlation for Hate Speech detection.

However, no study has been done for detecting social biases from toxic languages, which is a challenging task due to the conceptual overlap

between hate speech and social bias. Using a thorough guideline, we attempt to uncover harmful biases in toxic language datasets. The curated dataset is discussed in length in the next section, as are the definitions of each category label and the annotation procedure.

3 ToxicBias Dataset

We develop the manually annotated *ToxicBias* dataset to enable the algorithm to correctly identify social biases from a publicly available toxicity dataset. Below, we define social bias and the categories taken into account in our dataset. The comprehensive annotation process that we use for dataset acquisition is then covered.

3.1 Social Bias

People typically have preconceptions, stereotypes, and discrimination against other who do not belong to their social group. Positive and negative social bias refers to a preference for or against persons or groups based on their social identities (e.g., race, gender, etc.). Only the negative biases, however, have the capacity to harm target groups (Crawford, 2017). As a result, in our study, we *focus on identifying negative biases* in order to prevent harmful repercussions on targeted groups. Members of specific social groups (e.g., Women, Muslims, and Transgender individuals) are more likely to face prejudice as a result of living in a culture that does not sufficiently support fairness. In this work, we have considered five prevalent social biases:

- **Gender:** Favoritism towards one gender over other. It can be of the following types: Alpha, Beta or Sexism (Park et al., 2018).

- **Religion:** Bias against individuals on the basis of religion or religious belief. e.g. Christianity, Islam, Scientology etc (Muralidhar, 2021).

- **Race:** Favouritism for a group of people having common visible physical traits, common origins, language etc. It is related to dialect, color, appearance, regional or societal perception (Sap et al., 2019).

- **LGBTQ:** Prejudice towards LGBTQ community people. It can be due to societal perception or physical appearance.

- **Political:** Prejudice against/towards individuals on the basis of their political beliefs. For example: liberals, conservatives, etc.

Categories	Targets
Political	liberal, conservative, feminist, etc.
Religion	christian, jew, hindu, atheist, etc.
Gender	men, women
LGBTQ	gay, lesbian, homosexual, etc.
Race	black, white, asian, canadians, etc.

Table 1: Bias categories and corresponding targets.

For all of these categories, target terms are the communities towards which bias is targeted.

3.2 Social Bias *Vs* Hate Speech

While Social Bias and Hate Speech may appear the same at first look, they are not. The differences between them are quite subtle. While hate speech is always associated with negative sentiment, social bias can also have positive sentiments. Social bias is preconceived belief toward or against specific social identities, whereas hate speech is an explicit comment expressing hatred against an individual or a group. Not all hate speech is biased, and not all biased speech is hate speech. We will use the following examples to demonstrate the differences:

- Some comments are merely toxic without containing any social biases in them, e.g.
  IM FREEEEE!!!! WORST EXPERIENCE OF MY F**K-ING LIFE

- Toxic comments can be hate speech but not necessarily biased, e.g.
  you gotta be kidding. trump a Christian, nope, he is the devil, he hates blacks, Hispanics, muslims, gays, Asians, etc.

- Some comments are just biased with negating sentiment without containing any toxicity or hate speech in them, e.g.
  All Asian people are bad drivers.

3.3 Annotation Process

The dataset we used for annotation is collected from a Kaggle competition named "Jigsaw Unintended Bias in Toxicity Classification (jig, 2019; Research Data, 2018)". It has around two million Wikipedia comments annotated for toxicity. The data also has several other toxicity subtype attributes such as *severe toxicity, obscene, threat, insult, identity_attack, and sexual_explicit*.

We discovered that, with the exception of the *identity_attack* column, all of the columns in this dataset are redundant for the social bias detection

134

comment_text	label	category	target	implication
I wish gay men would just leave children alone.	bias	lgbtq	gays	gays are pedophiles
Islam, unfortunately, is a violent cult.	bias	religion	muslims	muslims are violent
Why does Netflix hate gay people?	neutral	-	-	-
Liberal women are so #king ugly!	bias	political, gender	liberals, women	prejudice against liberal women
Why are black thugs coming out here murdering locals? What is the real story? Please someone..hire an investigative journalist.	bias	race	blacks	prejudice against blacks

Table 2: Annotation of Bias along with its category, targeted groups and implication

task. *Identity_attack* shows denigration of the comment towards a identity. This identity could be related to race, religion, political affiliation, gender, or sexual orientation. So, we filtered out only those comments for which *identity_attack* values are greater than or equal to 0.5. We annotated this filtered dataset for the presence of social bias. We have considered only *five bias categories* for our annotation and *possible targets* listed in Table 1. We did not include other categories due to their low presence in the original dataset. The targets describe any social or demographic groups that is targeted in the comment. Bias implications are annotated in addition to bias categories and relevant targets. Table 2 shows a sample annotation of this filtered dataset. The bias implications are simple *free-text* reasons showing the stereotype towards the target group.

The final dataset contains 5409 cases with multiple label annotations. There are 120 distinct terms for target annotation divided into five categories. To check the consistency of our framework and to categorize biases, two different annotators annotated the data independently. Considering the complexity of the task, we provided a detailed guideline to each of the annotators. Following the thorough guidelines by Singh et al. (2022), we developed a series of questionnaires for each categories to assist the annotators. Inter-annotator agreement was assessed for the first 2500 occurrences, and a Cohen's Kappa value of 64.3 was found, indicating good agreement between annotators. The figure 2 depicts the distribution of data among multiple categories. All the disagreements between annotators were resolved by adjudication with the help of an expert. For details about the annotators, please refer A.2.

Out of 5409, our dataset has **4325** bias instances (80% of dataset) and **1084** neutral (not biased towards any identity). The number of instances for each category across train, dev., test are shown in Table 3.

Categories	train	dev	test	total
bias	3460	346	519	4325
neutral	867	86	131	1084
race	1769	181	252	2202
religion	1257	120	196	1573
gender	293	24	41	358
lgbtq	453	41	82	576
political	172	20	26	218

Table 3: Distribution of different categories across three splits of Train, Dev. and Test.

3.4 Annotation Challenges

Several issues were encountered while annotating harmful datasets, which must be handled to annotate and build a consistent dataset.

- Quoting someone else statements considered biased? Example:
  Trump said "Mexicans are rapists and drug dealers".

  One may argue that such statements are not biased as they are merely quoting someone else's opinions that are not theirs. However, we label them biased because quoting someone else's statement is just as detrimental and harmful.

- We believe that asking questions about an issue may not lead to bias. Example:
  Black idiot or white idiot. What is the difference?

  One would believe that this phrase is about idiots in general, without regard to race. As a result, there will be no prejudice. On the other hand, some may interpret that both blacks and whites are referred to as fools here.

- We also encountered statements lacking context. Example:
  Is that the white kind? I mean since you hate whites so much?

  Now here, we do not know whether the statement is talking about white colour or white

race. We label these kinds of sentences as neutral

- Some statements were purely made as a personal attack. These instances were labeled appropriately as biased or neutral. Example:
Trump pig latin. Oink, oink, oink, grab em by the poo say

  We label this statement as biased because Trump here represents a certain political party (community), unlike the below example:
settlers is a demeaning racist term. You Johnny are a white hating racist.

  Here Johnny is not a big political leader (name). So we mark this as neutral.

- We have encountered many sarcastic instances in the dataset and label them appropriately. Example:
Ah yes, re-education! That's what us nasty white folks need.

  We label this statement as neutral because it's a self-criticism sarcastically.

  Yeah - - because up until now, Islamic State really loved the US! And the West in general! They love us so much sometimes they cut off peoples heads to keep as a souvenir!

  The above statement was labeled as biased as it is sarcastically showing prejudice against Islam.

- Some statements are speaker dependent. Example:
Shit still happenin and no one is hearin about it, but niggas livin it everyday.

  This statement will not be biased if said by an African-American; however, it becomes highly offensive and biased if stated by someone else.

4 Experimental Setup

In this section we will discuss about different models trained for detection of social biases and their categories. For all our experiment, we split the data into train, development, and test (80:8:12) set. Since the dataset was imbalanced with respect to bias column, we split it in stratified manner.

4.1 Metrices

We report accuracy, macro F1-score, and AUC-based scores in accordance with best practice.
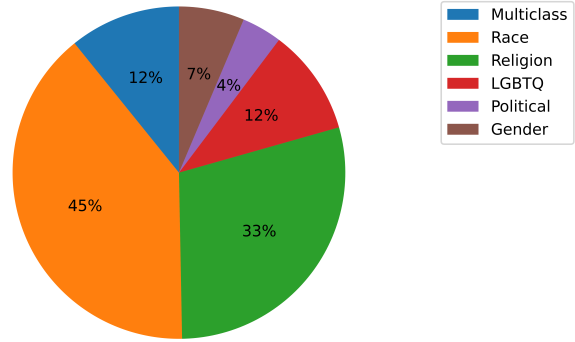

Figure 2: Distribution of bias categories in ToxicBias. It is observed that some instances qualified for multiple bias categories(12.22%)

These metrics would be used to assess the classifier's ability to distinguish between the bias and neutral texts along with bias categories. AUC stands for *Area under the ROC curve.* ROC curve depicts the tradeoff between true positive rate (TPR) and false positive rate (FPR). The AUC value is high when the TPR is high and the FPR is low.

Borkan et al. (2019) proposed AUC-based metrics to quantify the unintended model bias. These metrics compare the output distributions of instances that include the specific community word (subgroup distribution) with the rest (background distribution). The three AUC-based bias scores are as follows:

1. **Subgroup AUC (AUC_{sub}):** It calculates AUC exclusively on a subset of the data for a specified community word. A low score indicates that the model struggles to differentiate between bias and neutral comments related to the community word.

2. **Background Positive and Subgroup Negative AUC (AUC_{bpsn}):** AUC_{bpsn} uses the biased background instances and the neutral subgroup examples to determine AUC. A low score indicates that the model has high false positive rate. The model misinterprets neutral comments mentioning the community with biased comments missing it.

3. **Background Negative and Subgroup Positive AUC (AUC_{bnsp}):** It uses the neutral background instances and the biased subgroup examples to determine AUC. A low score suggests that the model has a high rate of false negatives. The model misunderstands biased comments that mention the community with neutral ones that do not.

136

	Model	P	R	F1	Acc
	Logistic Regression	0.67	0.50	0.46	0.84
Baselines	SVM	0.42	0.50	0.46	0.84
	Bi-LSTM + Glove	0.59	0.58	0.58	0.78
	BERT (Hierarchical)	0.62	0.66	**0.64**	**0.86**
Transformers w/o Aug	BERT (Multi-task)	0.90	0.52	0.49	0.81
	GPT2	0.62	0.66	0.62	0.71
	BERT (Hierarchical)	0.86	0.86	**0.86**	**0.88**
Transformers /w Aug	BERT (Multi-task)	0.86	0.86	**0.86**	0.87
	GPT2	0.81	0.86	0.84	0.81

Table 4: Performance of various models on bias detection task. We report results for baselines, and Transformer based training. For Transformer based training, we compare performances without data augmentation and with data augmentation. Best scores are shown in bold.

Model	Hierarchical				Multi-task			
	Acc	P	R	F1	Acc	P	R	F1
political	0.96	0.48	0.50	0.49	0.96	0.77	0.57	**0.61**
gender	0.95	0.47	0.50	0.49	0.95	0.84	0.71	**0.76**
race	0.84	0.81	0.83	0.82	0.86	0.86	0.88	**0.86**
religion	0.82	0.82	0.82	0.82	0.93	0.91	0.94	**0.92**
lgbtq	0.93	0.81	0.81	0.81	0.94	0.86	0.87	**0.86**

Table 5: Bias Category Detection Results. P, R, F1 and Acc are Precision, Recall, F1-score and Accuracy respectively. Best scores are shown in bold.

4.2 Baseline Models

We discuss several model architectures for detection of biases and their categories. For bias detection, which is a binary class classification task, we consider Logistic Regression (LR) with TF-IDF as our baseline model. Our baseline model gives 84% accuracy with 0.46 F1 score. The low F1 score clearly indicates that model has very high false positive rate and false negative rate. We also tried Support Vector Machine (linear kernel) with TF-IDF and LSTM (Huang et al., 2015) with Glove 300d word representation (Pennington et al., 2014). The best model is observed to be BERT (Devlin et al., 2019) with 0.64 F1 score. Two different model settings were used to detect biases and their categories. We will discuss each of them in detail in the following sections.

4.3 Hierarchical Model

In the hierarchical model, bias detection and category classification was done in two steps. Bias detection, a binary class classification task, is performed first. If the post has some biases, then its categories are detected next. Since a post may contain several biases, the bias category detection task was framed as multi-label classification. Bias detec-

tion results of several models in hierarchical model architecture is shown in Table 4. Bias category detection results in the hierarchical setting are shown in Table 5.

4.4 Multi-task Learning

In the context of classification, multi-Task Learning tries to improve the performance of numerous classification problems by learning them together. So instead of predicting bias and its category in two steps, we can train a model to predict them simultaneously in one step. Since there can be multiple biases in a post, we cannot use logistic regression or SVM in a multi-label classification task. Hence in this model architecture, we try LSTM and BERT models only. We use LSTM with a single output layer. The last dense layer of the LSTM comprises six neurons, one to detect bias and the other five to identify bias categories.

Precision (P), recall (R), F1 (macro values for all), and accuracy (Acc) for bias detection experiments in Multi-task architecture is shown in Table 4. Table 5 shows the comparison between hierarchical and multi-task model for category detection task.

4.5 Generation Framework

Considering the efficacy of GPT (Radford and Narasimhan, 2018) based model for classification, conditional generation tasks (Sap et al., 2020), we frame the prediction of categorical variables and implications as generation task. The input is a sequence of tokens as in Equation 1, where w_i are the tokens corresponding to comment text and [BOS], [SEP], [EOS] are start token, separator token and end token respectively. Two task specific tokens ([BON], [BOFF]) were added to the token

comment_text	Ground truth label	Predicted label	Ground truth category	Predicted category
Quran is a holy book.	neutral	bias	-	religion
So then I was all like "I'd rather get the black plague and lock myself in an iron maiden than go out with you.	neutral	bias	-	race
Do they come in men's sizes?	neutral	bias	-	gender
What I've just shown is that this happens in every black hole.	neutral	bias	-	race

Table 6: Error analysis showing model biases from predictions of Multi-task BERT model without augmentation.

Variables	BLEU-2	RougeL
Categories	61.60±0.96	88.23±1.23
Target subgroup	52.95±2.84	77.58±4.21
Implications	33.4±1.55	39.5±1.20

Table 7: Evaluation of various generation tasks. The standard deviations for three runs are also reported.

vocabulary which were used as $w_{[\text{bias}]}$ in the input. Here, [BON], [BOFF] correspond to bias and neutral instances respectively. As we have many inputs with multiple bias categories and targets, we combine them using a comma separator in the raw text. While encoding the input we use $w_{[\text{C}]_i}, w_{[\text{T}]_i}$ as the token corresponding to them respectively. Similarly, $w_{[\text{R}]_i}$ is used for representing the tokens corresponding to implications.

$$\mathbf{x} = \{[\text{BOS}], w_i, [\text{SEP}]\ w_{[\text{bias}]}, [\text{SEP}]\ w_{[\text{C}]_i}, [\text{SEP}]w_{[\text{T}]_i}, [\text{SEP}]w_{[\text{R}]_i}, [\text{EOS}]\} \tag{1}$$

For this experiment, we finetune the GPT-2 (Radford et al., 2018) model with commonly used hyperparameters. For training we use cross-entropy loss as cost function. During inference, we first calculate the normalized probability of $w_{[\text{bias}]}$ conditioned on the initial part of input and then append the highest probable token to the input and generate rest of the tokens till [EOS].

We use BLEU-2 (Papineni et al., 2002) and RougeL (Fmeasure) (Lin, 2004) as the metrics to calculate the performance of the model for category, target and implication of the comment text(Table 7) and macro F1 as metric for bias evaluation(Table 4). Performance for category generation is better than other two variable as it has less ambiguity whereas the low performance for implications show the variability in the annotation for implications.

We report all the hyperparameters and training setup in appendix A.3.

4.6 Results and Discussion

From the above tables, we can infer that the BERT-based hierarchical model outperforms Multi-task and GPT-2 models for the bias detection task. In addition, category identification is performed more effectively by multi-task model. This is quite apparent from the fact that in multi-task learning each model aid each others due to shared parameters. We also see that F1 scores for political and gender categories are less when compared to other categories in the category detection task. One plausible explanation might be the low frequency of such biases in our dataset. The BLEU scores of category and target subgroup generation are higher than those of bias implication generation, as shown in Table 7. The fundamental reason for this is that the bias category and target have less distinct bigrams/tri-grams than the implications. We show an error analysis for categorical prediction tasks after training the models. We discovered that the model predicts incorrectly for some simple sentences. A few examples are also listed in the Table 6. This most likely occurred due to lexical overfitting/ model bias of the trained model towards some of the identity specific terminologies. In the next section, we discuss an approach to reduce this unwanted behavior of the model using counternarrative augmentation.

5 Mitigation of Model Bias

When we look at the incorrectly classified comments in Table 6, we observe that they contain community words such as 'blacks,' 'Quran,' and so on. Sometimes, due to the presence of these community terms, our model predicts that these comments would be biased. In essence, our initial model is latching onto some community related terms and hence suffers from model bias. According to (Zueva et al., 2020), most existing models provide predictions with certain bias. Even if the statement itself is not toxic, the model commonly classifies it as toxic if it includes specific frequently

comment_text	Ground truth category	Prediction (**without** augmentation)	Prediction (**with** augmentation)
Quran is a holy book.	none	religion	none
Islam is a peaceful religion.	none	religion	none
Muslims are invaders.	religion	religion	religion
Do they come in men's sizes?	none	gender	none
What I've just shown is that this happens in every black hole.	none	race	none
I'm sick of niggers, it's a bad neighborhood. I don't like being around them, they're animals.	race	race	race

Table 8: Error analysis showing the effectiveness of counter-narrative augmentation. After training with augmented data, the resulting model does not latch on to the community related terms.

Subgroup	Count	% Bias
all comments	5409	79.92
black	1291	79.55
jewish	269	74.34
lgbt	778	77.24
muslim	1263	87.01
female	586	76.45

Table 9: Percentage of bias comments by identity terms such as black, jewish, lgbt, muslim, female in the *ToxicBias* dataset.

targeted identities (such as women, blacks, or Jews). Similarly, our model incorrectly labels comments referencing particular identities, such as Blacks, Muslims, and Whites, as social bias. Model biases emerge when identity words like Blacks, Whites, and Muslims appear more frequently in biased comments than in neutral comments. If the training data for a machine learning model is skewed towards certain terms, the final model is likely to acquire this bias. Table 9 shows the bias percentage in *ToxicBias* for several identities/subgroups, indicating the imbalance for bias labels among those identities and emphasising the importance of AUC-based metrics resilient to these data skews.

Counter-narratives: Despite enormous attempts to build suitable legal and regulatory responses to hate content on social media platforms, dealing with hatred online remains challenging. If hate speech is addressed with standard content deletion or user suspension methods, censorship may be accused. Actively addressing hate material through counter-narratives (i.e., informed textual responses) is one potential technique that has received little attention in the academic community thus far. A counter-narrative (also known as a counter-comment or counter-speech) is a reply that provides non-negative feedback through fact-based arguments and is often recognized as the

most effective way to deal with hate speech.

subgroup	$AUC_{sub}\uparrow$	$AUC_{bpsn}\uparrow$	$AUC_{bnsp}\uparrow$
black	0.48	0.50	0.49
jewish	0.47	0.50	0.49
lgbt	0.81	0.83	0.82
muslim	0.82	0.82	0.82
female	0.81	0.81	0.81

Table 10: AUC based scores for subgroups on bias detection model trained without data augmentation. Higher AUC values for each target subgroup indicate reduced lexical overfitting/model bias for those targets.

subgroup	$AUC_{sub}\uparrow$	$AUC_{bpsn}\uparrow$	$AUC_{bnsp}\uparrow$
black	0.86	0.78	0.97
jewish	0.91	0.93	0.91
lgbt	0.89	0.91	0.93
muslim	0.96	0.97	0.86
female	0.93	0.94	0.93

Table 11: AUC based scores on bias detection model trained after data augmentation. Higher AUC values for each target subgroup indicate reduced lexical overfitting/ model bias for those targets.

We use two counter-narrative datasets to reduce the model biases: CONAN (Chung et al., 2019) and Multi-target CONAN (Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco, 2021). These datasets provide counter-narratives to hate speech or stereotypes directed towards social groups such as Muslims, Blacks, Women, Jews, and LGBT people. So they do not contain any negative social biases towards those groups. Combining these counter narratives ensures that the resulting dataset will have more neutral/positive instances mentioning those identity terms. Adding these counter narratives to our dataset significantly decreased model biases. We used total of 7219 counter-narratives related to jews (593), muslim (4996), black (352), homosex-

ual_gay_or_lesbian (617), and female (661). As illustrated in table 10, black and jewish identities suffer from both high false positives and high false negatives. However, after counter-narrative augmentation, the resulting model appears to be capable of dealing with the problem of model bias. Table 11 shows the reduction in model bias using AUC-based metrics. Table 8 includes an error analysis to show how CONAN has helped reduce model bias.

6 Conclusion and Future Work

We have demonstrated that identity attacks or hate speech often incorporate social biases or stereotypes. However, not all hate speech can be labeled as social bias. Some of them are merely personal insults. Filtering out such biases from hate speech is not a trivial task. Furthermore, we have frequently observed that detecting bias without context for the comment or demographic information of the comment holder makes the annotation much more challenging. However, detecting these social biases from toxic datasets, which are available in relatively large amounts, will be a useful starting point for social bias research in other forms of text.

The issue of model bias is also observed during inference. The imbalanced existence of particular community terms (muslims, whites, etc.) might lead to a model labeling a comment as biased. To attenuate model biases, we used counter-narratives and showed that they help significantly to reduce model biases. From our study, we also observe that biases can have directions too. So basically, biases can occur against specific communities and in favour of a community. We intend to detect such biases in future work.

7 Acknowledgements

8 Limitations

The most notable limitation of our work is the lack of external context and small-sized dataset. In our present models, we have not considered any external context that can be useful for the categorization task, such as the profile bio, user gender, post history, etc. Our work currently considers only five types of social biases, not all other possible dimensions of bias. We also concentrated on using only the English language in our work, and the dataset is oriented toward western culture. The bias annotations in the dataset may not be very relevant to people of non-western culture. Furthermore, Multilingual bias is not taken into account.

References

2019. Jigsaw unintended bias in toxicity classification.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification.

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

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51:1–30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189:274–281. AI in Computational Linguistics.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

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

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *CoRR*, abs/2008.06460.

Deepa Muralidhar. 2021. *Examining Religion Bias in AI Text Generators*, page 273–274. Association for Computing Machinery, New York, NY, USA.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Marie Plaisime, Candace Robertson-James, Lidyvez Mejia, Ana Núñez, Judith Wolf, and Serita Reels. 2020. Social media and teens: A needs assessment exploring the potential role of social media in promoting health. *Social Media + Society*, 6(1):2056305119886025.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Civil Research Data. 2018. Civil comments.

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

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. *ACL*.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi Sultan, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5274–5285, Marseille, France. European Language Resources Association.

Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80.

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*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in Russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69, Online. Association for Computational Linguistics.

A Appendix

A.1 Ethical Considerations

Our work aims at capturing various social biases in toxic social media posts and demonstrates the annotation quality on biases in one of existing dataset. We also discuss the challenges we faced while doing the annotation of the dataset, specifically due to the absence of context for each instance in the dataset. Also, study of social biases come with ethical concerns of risks in deployment (Ullmann and Tomalin, 2020). As these toxic posts can create potentially harm to any user or community, it is required to conduct this kind of research to detect them. If done with precautions, such research can be quite helpful in automatic flagging of toxic and harmful online contents.

Researchers working the problem of social bias detection on any form of text would benefit from the dataset we have collated and from the inferences we got from multiple training strategies.

A.2 Annotator Demographics and Treatment

Both the annotators were trained and selected through extensive one-on-one discussions, and were working voluntarily. Both of them went through few days of initial training where they would annotate many examples which would then be validated by an expert and were communicated properly about any wrong annotations during training. As there are potential negative side effects of annotating such toxic comments, we used to have regular discussion sessions with them to make sure

they are not excessively exposed to the harmful contents. Both the annotators were Asian male and were of age between 23 to 26. The expert was an Asian female with post-graduation degree in sociology.

A.3 Training Details

A.3.1 BERT Training

We finetune 12 layer BERT base uncased with batch size of 32 for two epochs. Max token length of 128 is used. We experiment with learning rates of $2e-5, 3e-5, 4e-5, 5e-5$ with AdamW(Loshchilov and Hutter, 2019) optimizer and epochs of $5, 10, 20$. We also use a dropout layer in our model. AdamW optimizer with learning rate = $5e-05$, epsilon = $1e-08$, decay = 0.01, clipnorm = 1.0 were used.

A.3.2 GPT-2 Training

We finetune GPT-2 with a training batch size of 1, gradient accumulation step as 4, and 200 warm up steps. Experiments were run with a single GeForce RTX 2080 Ti GPU. Finetuning one GPT-2 model took around 40 minutes for 5 epochs.

We have kept all the parameters of BERT and GPT-2 trainable. All of our implementations uses Huggingface's transformer library (Wolf et al., 2020).