ADS Audit: Unintended Bias in Toxicity Classification

Chloe Zheng, Preston Harry

DS-GA 1017 Spring 2022

1 Introduction

The ADS considered in this report is an entry to the Kaggle competition "Unintended Bias in Toxicity Classification" sponsored by Jigsaw (Borkan 2019). The aim of this competition is to predict the "toxicity" of comments made on the Civil Comments platform–an application to allow users to comment on and discuss articles from independent news sites while regulating the toxicity of those comments. "Toxic" comments are defined as anything "rude, disrespectful or otherwise likely to make someone leave a discussion." The ADS should flag toxic comments to regulate online conversations and ideally protect voices that are discouraged from participating online. Past models have had difficulty with predicting toxicity pertaining to one's identity or other protected characteristics by associating any mention of the protected characteristic with toxicity. Any ADS should both identify toxic comments while mitigating harmful bias related to users' identities.

Simply maximizing accuracy may exacerbate biases by disproportionately assigning toxic labels to certain groups, causing disparate treatment in comment regulation. Meanwhile, guaranteeing an equitable distribution of toxicity classification may reduce prediction accuracy. The challenge is to reasonably balance fairness and accuracy such that the ADS that does not disparately impact those with protected characteristics while still retaining a high enough accuracy to be useful.

2 Data

2.1 Data Collection

This ADS uses data containing all public comments from the Civil Comments application from its opening in 2015 to its closing in 2017. These comments total to about 1.8 million. Jigsaw, a Google subsidiary focused on understanding technology's role in social issues, used annotators to manually label comments as severe, obscene, an identity attack, insulting, threatening and/or sexually explicit. The final dataset includes a target indicating whether a comment is toxic, the text of the comment, labels to describe the type of toxicity, and a series of comment topic indicators including race, gender, sexuality, religion, and disability.

These comment topic indicators are not meant to be used for modelling, but to allow for bias analysis on these protected attributes. Values for these identity features are derived from multiple annotators indicating whether a comment contains information about a given identity. Comments are given a score between 0 and 1 for each identity equal to the fraction of annotators that believe the comment pertains to that identity.

However, only select identity fields are fully represented across all comments in the data. Identities with at least 500 positive examples in Kaggle's test set are fully represented in the training data with no missing values. The annotator scores for these identities are mapped to a boolean vector at some unspecified threshold. Original annotator scores are not retained in the data, so this threshold cannot be derived. These sufficiently frequent identity groups are: male, female, homosexual, Christian, Jewish, Muslim, Black, White, or having mental illness. The Kaggle competition only considers these listed subgroups when evaluating models and unintended bias.

Remaining infrequent identity fields which have less than 500 positive examples in the test set retain their annotator scores and are not mapped to a boolean vector. However, for approximately 80% of observations, annotator scores for these infrequent identity fields are replaced with missing values. It is unclear why this is the case. If it is meant to offer some form of security regarding the annotator's labels, it seems like an ineffective option as each observation still contains the comment's text. It is also unclear if the values were replaced with missing values at random or according to some rule.

2.2 ADS: Input Data

The input of the model only has one feature "comment text," where each observation represents one comment from a given user.

Each comment is converted to a numeric vector where each element of the vector represents each word in the comment. Each comment vector is also padded with zeros so that each comment vector is 250 in length to account for differing sentence length. For example, given a comment with n tokens, the corresponding comment vector is a one-dimensional vector of length 250, such that there are 250-n leading zeros, and the last n values are the sequence of numbers representing each word of the comment. There are no missing or empty comments in the sample. Since there is effectively only one feature being input to the model, it is not relevant to calculate pairwise correlations of the input data.

2.3 ADS: Output Data

For a given comment, the output of the model is the probability that the comment is toxic. Given some threshold, if the probability is above the threshold the comment is toxic, otherwise it is not. The developers do not indicate a threshold, so we assume the threshold is 0.5.

3 Implementation and Validation

Given computational limitations, only 10% of data is used to verify pre-processing, training, and evaluation code.

3.1 Data Pre-processing

Text pre-processing is relatively straightforward. The authors use a Keras tokenizer to set everything to lowercase and remove punctuation. Removing capitalization could introduce bias in this context because intuitively, capitalization may suggest aggressive and toxic tone in comments. Other conventional text preprocessing, such as removing stop words or stemming, are not performed.

3.2 Model Implementation

The developers implement a Keras Convolutional Neural Network that consists of one embedding layer, three layers of 1D convolutions and max pooling, and one dense layer. An embeddings matrix is constructed using Stanford University's pre-trained 100 dimensional text vectors, which was trained on 6B tokens and 400K vocabulary from the 2014 Wikipedia and Gigaword 5 dataset (Pennington et. al, 2014). The 3 layers of convolutions and max pooling utilizes the ReLU activation function and same padding, and the final dense layer has a softmax activation function for final probability predictions. The model uses categorical cross entropy loss as the objective. Hyperparameters are also set as follows: dropout_rate=0.3, learning_rate=0.00005, num_epochs=10, and batch size=128.

3.3 Goals and Validation

The Kaggle competition's stated goal was "to build a model that recognizes toxicity and minimizes ... unintended bias with respect to mentions of identities."

3.3.1 Accuracy

The developers generally satisfy the first goal, maximizing auc scores and achieving high accuracy in classifying toxicity on the validation dataset. We verify a 93% validation accuracy using 10% of the total data and the same 80/20 training and validation split.

3.3.2 Bias Analysis

The identity indicator features provided by Jigsaw allows for bias analysis on the validation data. However, Jigsaw has only provided complete data for identities with at least 500 positive examples in Kaggle's test set. These frequent identities are binary indicators and have values for every comment.

Since Kaggle is only concerned with evaluation on the frequent identity fields, this ADS considers only those fields. Unfortunately, Kaggle does not give clear direction on goals and expectations for bias evaluation. To analyze bias, the authors maximize auc scores for subgroups, and for "BPSN", and "BNSP" (Table 1). BPSN is the AUC of the within-subgroup negative examples and the background positive examples, and BNSP is the AUC of the within-subgroup positive examples and the background negative examples. The existence of these "bias metrics" might have satisfied the competition's stated goals, but overall, it is not clear what the goal and conclusion is from these metrics.

To evaluate bias and fairness, it must be clear what differences in a metric across subgroups reveals about subgroup treatment and how results indicate whether unprivileged and privileged subgroups are treated fairly relative to one

Table 1: AUC for Frequent Identity Identifiers

Subgroup	Subgroup Size	Subgroup AUC	BPSN AUC	BNSP AUC
white	534	0.745796	0.738832	0.922289
black	315	0.748359	0.757065	0.907939
homosexual_gay_or_lesbian	227	0.781783	0.830528	0.878998
muslim	447	0.825704	0.773116	0.943594
psychiatric_or_mental_illness	98	0.826577	0.908092	0.826843
male	873	0.836382	0.803792	0.932160
female	1052	0.855858	0.823472	0.925211
jewish	139	0.881113	0.827358	0.943589
christian	805	0.886281	0.881122	0.907277

another. The authors fail to explain consequences for a lower or higher AUC per subgroup, and they fail to define reasonable privileged and unprivileged groups or indicate when groups of metrics exhibit fairness.

For instance, they could have defined "White" as privileged, "Black" as unprivileged, and showed how subgroup AUC scores for these groups are comparable, indicating there is not a disparity between how the model successfully labels these comments. For religion related attributes, "Muslim", "Christian", and "Jewish" AUC scores are not comparable, which could indicate bias. However, AUC does not tell us exactly how a subgroup is being incorrectly labeled. Finally, features "homosexual_gay_or_lesbian" and "psychiatric_or_mental_illness" have no reasonable corresponding privileged subgroups. Therefore, it is not possible to analyze group fairness with their reported AUC scores.

4 Outcomes

While this ADS only considers the same identity fields that Kaggle is concerned with, we extend their analysis of subgroup AUC to all subgroups, accepting potential uncertainty and increased variance for these largely missing fields. Additionally we consider differences in false positive rates and false negative rates across subgroups.

4.1 Unpopulated Identity Features

To analyze all subgroups, we must first account for the infrequent identity fields for which approximately 80% of observations have missing values. These identity fields are still represented as the fraction of annotators who believed a particular identity was mentioned within a comment. Kaggle does not specify Jigsaw's method for identifying a proper threshold to determine if a comment is considered to have mentioned a given identity. We then use the mean annotator score in the training set for each identity field as the cutoff.

This method is certainly imperfect as it introduces quite a bit of uncertainty into each identity field. Using the population mean results in lower annotator scores for a given identity counting towards positive identification of that identity. Using a cutoff of 0.5 to ensure that a majority of annotators agree is an option, but most annotator scores are lower than 0.5 and this results in too few observations within subgroups to provide any useful analysis. Instead, we choose to analyze subgroups while understanding that they may not perfectly represent each identity.

We additionally ignore the missing values for each identity field–which is equivalent to treating those observations as being unrelated to those identity fields. While any evaluation metric calculated for observed identity subgroups should be an unbiased estimator for the same evaluation metric on the entire validation set, with almost 80% of values missing, these evaluation metrics have very high variance. However, this unbiased estimation assumption only holds if the missing identity fields are missing at random. We have no way to identify why annotator scores for these comments were excluded in particular. If they were not excluded at random, it further calls into question the validity of metrics calculated on observed subgroups.

4.2 Fairness Evaluation

4.2.1 AUC

Despite the issues with missing values for some subgroups, we still attempt to expand the previous AUC analysis to include all subgroups. Results are grouped by identity category: race, gender, sexuality, religion, and disability.

Table 2: AUC for Frequent and Infrequent Identity Identifiers

Category	Subgroup	Count	Subgroup AUC	BPSN AUC	BNSP AUC
Race	Black	315	0.726531	0.718669	0.92483
	White	534	0.733472	0.695652	0.941151
	Latino	130	0.819444	0.831462	0.900408
	Asian	224	0.855477	0.873304	0.890118
	Other race or ethnicity	396	0.858491	0.837326	0.922001
Gender	Other gender	53	0.790404	0.818305	0.910186
	Male	873	0.832295	0.798342	0.928129
	Female	1052	0.838212	0.815186	0.922816
	Transgender	131	0.861997	0.838199	0.918891
Sexuality	Homosexual gay or lesbian	227	0.763163	0.833075	0.873226
	Heterosexual	74	0.829301	0.837348	0.900779
	Bisexual	67	0.853276	0.818867	0.929534
	Other sexual orientation	83	0.857276	0.848385	0.913116
Religion	Atheist	46	0.754167	0.754966	0.913059
	Buddhist	34	0.828125	0.883227	0.855853
	Muslim	447	0.831754	0.755824	0.950785
	Jewish	139	0.837999	0.806968	0.919018
	Other religion	322	0.861872	0.846039	0.919025
	Christian	805	0.888592	0.869347	0.916999
	Hindu	33	0.95679	0.820656	0.977902
Disability	Psychiatric or mental illness	98	0.775338	0.882188	0.820042
	Other disability	59	0.785714	0.860255	0.845204
	Physical disability	59	0.821078	0.880866	0.847961
	Intellectual or learning disability	49	0.921569	0.893824	0.922545

The inclusion of all subgroups allows for broader comparison within categories (Table 2). Whereas the previous analysis suggests the model has comparable AUC across races, including more racial subgroups reveals that the model actually returns worse AUC scores for Black- and White-related comments compared to other races. Similar trends exist for gender and religion categories. However, note the wide gap in AUC between Atheist- and Hindu- related comments may be due to variance from the small number of observations.

For sexuality and disability categories, there now exist subgroups to compare against to better assess fairness. The model has more trouble predicting for comments related to homosexuality compared to comments related to other sexualities. Within the disability category, the model performs quite well regarding learning disabilities, but struggles with mental illnesses and other disabilities in comparison.

These AUC values alone don't mean the model must necessarily be unfair. Ideally, a model has similar predictive power across different attributes, but it is unclear how the model may be unfair with just this information. Understanding how the model misclassifies examples within different subgroups offers more insight.

4.2.2 FPR and FNR

We evaluate False Positive Rates (FPR) and False Negative Rates (FNR) for all subgroups to assess group fairness. High FPR for one subgroup relative to another may suggest the model associates that subgroup with toxicity. High FNR for a subgroup may imply the model associates that subgroup with non-toxicity. Results are also grouped by the following attributes: race, religion, sexuality, gender, and disability. This allows us to check for disparate treatment based on protected attributes.

To calculate stable metrics across subgroups, we train models on 10 bootstraps sampled with replacement from the training data, evaluate on the same validation dataset for each model, and plot boxplots grouped by attributes. Figure 1a and Figure 1b show all boxplots for all identity categories.

Overall, FPR is low and FNR is high across subgroups and attributes. This could be a consequence of a low base rate of toxic comments in the training dataset - about 8% are labeled toxic. However, lack of toxic comments in the dataset could result from pre-existing bias, reflecting annotators' more lenient opinions towards comments.

Consider a stakeholder, a vendor, that is looking to use this model to regulate user content on a social media platform. A more conservative approach to labeling comments as toxic could be a beneficial bias to users if comments are truly not

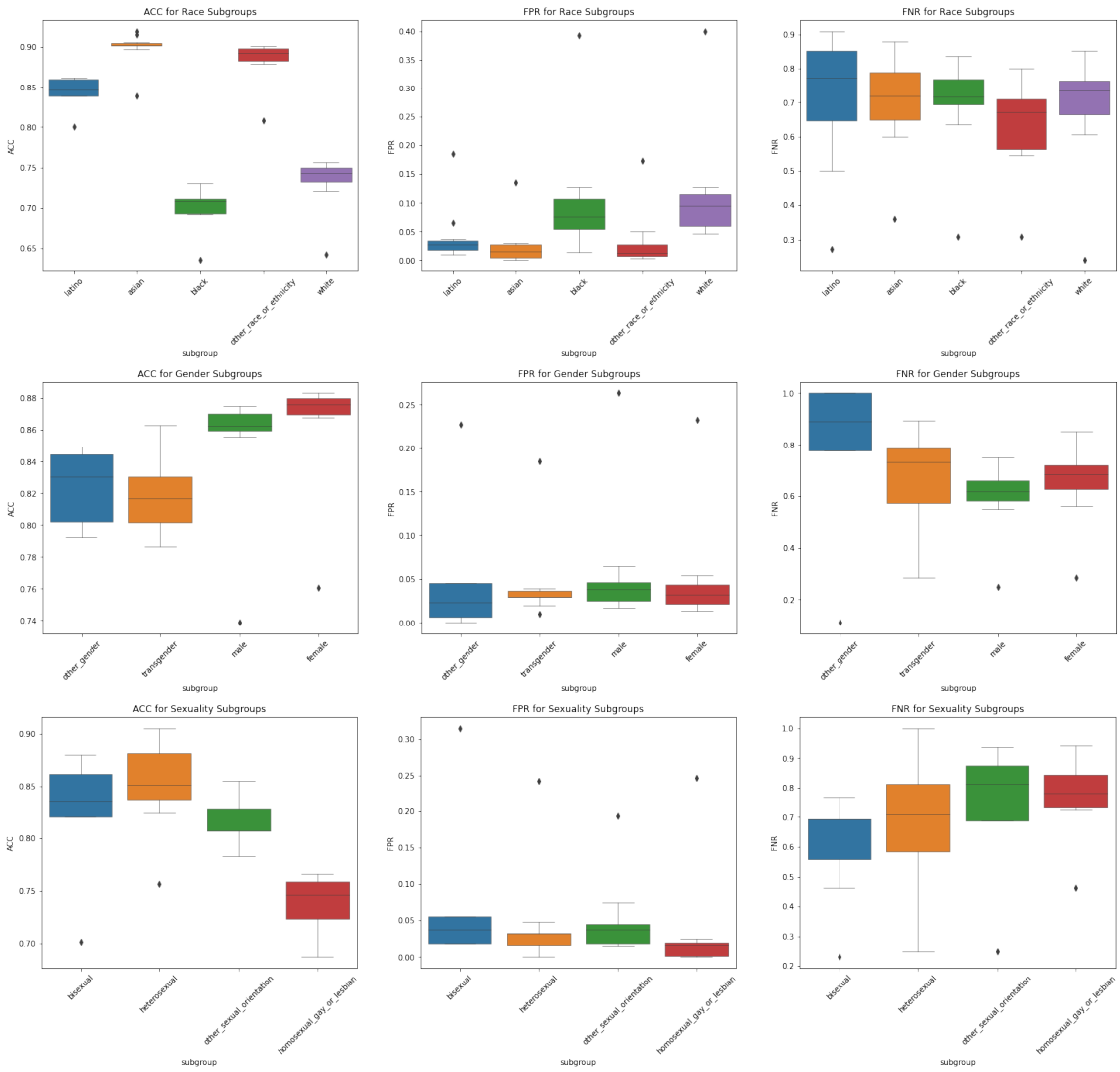


Figure 1a: Accuracy, FPR, and FNR results from 10 trials; subgroup categories: race, gender, sexuality

"inappropriate" the majority of the time, or users are interested in maintaining freedom of speech. Alternatively, being more lenient towards potentially toxic comments could be problematic if the users are interested in reliable moderation of toxicity. For instance, sexuality and gender subgroups have similarly low FPR and high FNR. Low FPR means that non-toxic comments related to these protected features are labeled not toxic, which suggests the model does not associate protected features with toxicity. This seems appropriate because these features can be used with neutral or positive intentions. However, high FNR means toxic comments that include these protected features are also often labeled not toxic. This suggests that the model is not successfully labeling toxic comments and not meeting one stated goal of identifying toxic comments.

A notable result is that comments containing White and Black identifiers have almost double the FPR of other racial subgroups. Considering historic and current polarization between White and Black people in the U.S at the time this data was collected, particularly the 2016 election, shootings, and other major events, the higher FPR seems to be a result of pre-existing bias. This could be problematic because the model might be associating racial identification with toxicity.

Note that the results also have significant variation due to the construction of the validation data. Many of the "other" categories, and subgroups in "sexuality", "gender", and "disabilities" were not frequently identified in comments. This could be due to annotator bias or a general lack of diversity in data. Therefore, we cannot make significant conclusions for many of the subgroups or gain insights for additional intersectional subgroups.

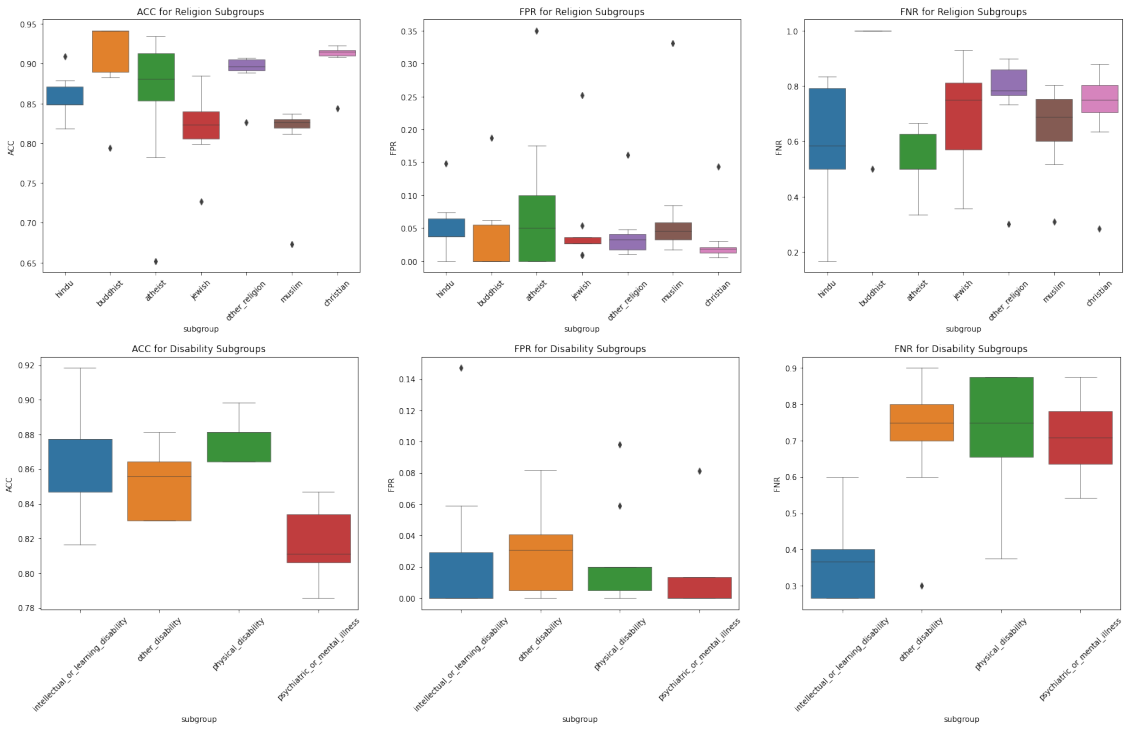


Figure 1b: Accuracy, FPR, and FNR results from 10 trials; subgroup categories: religion, disability

4.2.3 SHAP Explanations

To further assess fairness, we investigate the SHAP values of comments including racial features Black and White. We chose these two categories because of the high volume of relevant comments and they have higher FPR compared to other racial subgroups. SHAP explanations can verify if the ADS unreasonably associates these features with toxicity.

We fit a SHAP kernel explainer on 100 randomly sampled training observations. With that explainer we estimate SHAP values for 150 randomly sampled validation observations for both Black-related and White-related comments. The following tables display the top 5 and bottom 5 average SHAP values, for each set of comments. The tables also include the number of instances for which each word is evaluated positively and negatively by the SHAP explainer. Note that since comments may use terms multiple times, the actual term count may sum to a number greater than 150.

Table 3 includes SHAP values for comments that mention "black" (left) and "white" (right). The results demonstrate that tokens like "idiot," "stupid," "shit," and "bigot" have the most positive weight, indicating toxicity of comments; while tokens like "wp," "text," "duh," and "village" have negative weight, indicating non-toxicity of comments. These

Table 3: Top and Bottom 5 Words by Mean SHAP Values (All Words)

Black-related Comments					White-related Comments				
Word	SHAP Mean	SHAP Std. Dev	Positive Count	Negative Count	Word	SHAP Mean	SHAP Std. Dev	Positive Count	Negative Count
idiot	0.656609	0.012035	4	0	shit	0.359153	0.000000	1	0
stupid	0.517934	0.000000	1	0	bigot	0.182565	0.023506	2	0
crap	0.390789	0.000000	1	0	dumb	0.169451	0.000000	1	0
bigot	0.213302	0.000000	1	0	spineless	0.168221	0.000000	1	0
irresponsible	0.184091	0.000000	1	0	crazy	0.163414	0.003105	2	0
trailer	-0.035634	0.000000	0	1	deleted	-0.044967	0.000000	0	1
skewed	-0.037295	0.000000	0	1	funny	-0.048490	0.008651	0	2
immigration	-0.041773	0.000000	0	1	wikipedia	-0.051707	0.000000	0	1
text	-0.050299	0.000000	0	1	village	-0.052740	0.000000	0	1
wp	-0.060444	0.000405	0	2	duh	-0.092301	0.000000	0	1

Table 4: Top and Bottom 5 Words by Mean SHAP Values (Words in Top 10 Percentile of Use)

Black-related comments					White-related comments				
Word	SHAP Mean	SHAP Std. Dev	Positive Count	Negative Count	Word	SHAP Mean	SHAP Std. Dev	Positive Count	Negative Count
racist	0.046228	0.024620	21	0	racist	0.042397	0.031593	22	0
black	0.044818	0.037996	191	2	black	0.036626	0.028203	37	2
cops	0.042554	0.022657	10	0	white	0.036239	0.029082	181	4
guy	0.034466	0.021638	10	0	supremacists	0.031421	0.019892	15	0
white	0.029545	0.022709	101	4	women	0.024868	0.021751	16	1
professor	-0.004356	0.005762	0	9	also	-0.005441	0.008518	0	14
against	-0.004505	0.008147	3	26	population	-0.005863	0.010016	0	16
point	-0.004546	0.005724	0	9	which	-0.006734	0.007501	1	12
house	-0.005107	0.004528	0	9	comment	-0.011282	0.015161	0	10
players	-0.006255	0.008457	1	11	farm	-0.019280	0.014994	0	8

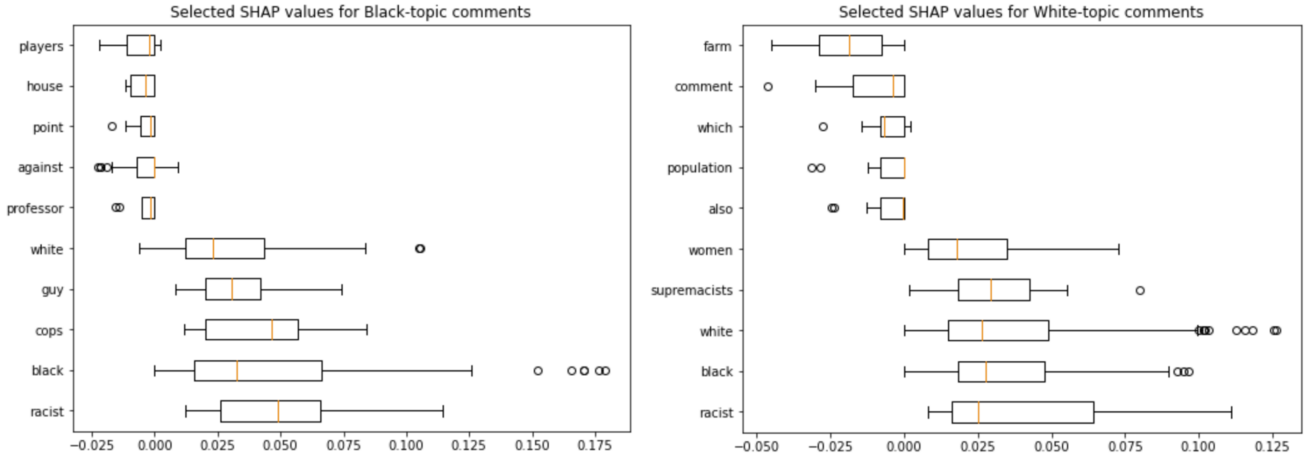
values seem reasonable given the negative connotation of the top 5 tokens and neutral connotation of the bottom 5 tokens for both subsets of comments.

However, for both sets of comments, the words with the highest and lowest average values are extremely infrequent in the data, with only one or two instances of use in comments. Table 4 shows terms that are in the top 10 percentile of frequent use. For Black-related comments, words that have more than 8 instances are shown while for White-related comments, words that have more than 7 instances are shown.

In both cases, identity related terms like “black,” and “white,” and even “women” have positive SHAP values, meaning the ADS associates these terms with toxicity. Additionally, terms like “racist,” “supremacists,” and “cops” are likely associated with racial topics. Furthermore, a large majority of instances of these terms are assigned a positive SHAP value. For Black-related comments, about 99% and 96% of instances of “black” and “white” contribute to a positive toxicity score, respectively. For White-related comments, about 95% and 98% of instances of “black” and “white” contribute to a positive toxicity score.

Figure 2 shows the mean and standard deviation for the top 5 and bottom 5 terms listed in Table 4. As with the tables, identity-related terms have positive average SHAP values with skew towards higher SHAP values.

Figure 2: SHAP Boxplots of Black and White related comments



Associating identity related terms with toxicity so prevalently is an unfair association that may be harmful to unprivileged groups. Users that simply identify with some identity in a comment may be flagged as toxic, disallowing commenters the freedom to discuss unprivileged groups or even their own identities.

To provide some examples of the consequences of associating identity-related terms with toxicity, consider the following two comments and their associated SHAP values. Figure 3 displays a toxic comment that has been falsely labeled as non-toxic by the ADS. Though one would expect swear words to be the largest determiner of toxicity here, ”black” has the largest positive SHAP value. The model fails to pick up on certainly toxic terms and spuriously treats ”black” as toxic. Figure 4 is a non-toxic comment that has been falsely labeled as toxic by the ADS. The phrase ”black and white” is a

common neutral expression and does not even pertain to race. Even so, the most important terms in incorrectly predicting a positive label are "white" and "black."

Figure 3: "Poverty increased, food stamp increases, debts and deficits increased greatly, black killing blacks increased exponentially in certain areas, auto debt record highs, student debt record highs, Middle East conflict never stopped. The saddest part is he got a peace prize for jack sh&&. That's obama basically."

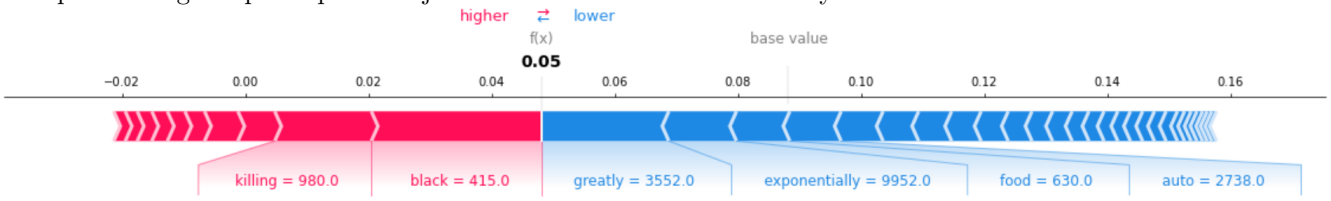
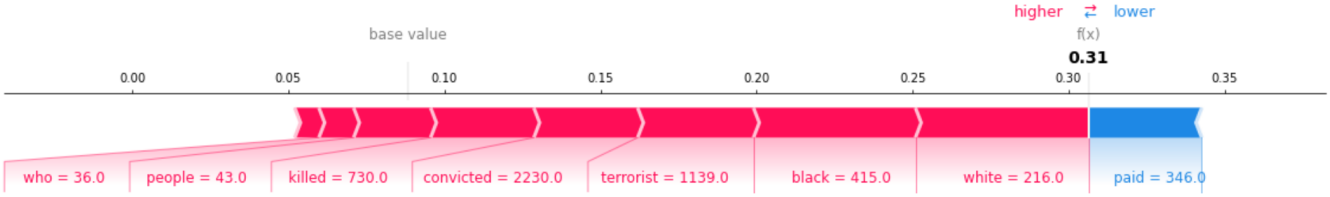


Figure 4: "It's pretty black and white, convicted and known terrorist who killed people gets paid $10.5M for it."



Ultimately, this ADS appears to directly associate identity-related terms to toxicity. This could be extremely detrimental to commenters who wish to identify a certain way or discuss identity on a platform that uses this ADS for moderation.

5 Summary

5.1 Data Suitability

The stated goal of this ADS is twofold: 1) to build a model that identifies toxic comments and 2) reduces unintended bias resulting from mentioning identity attributes.

These data are well-suited to the first goal–there are comments and toxicity labels. The quality of the toxicity labels may be questionable as the labels are subject to potential bias of the commenters. However, labels for something so subjective are difficult to come by, and identifying toxicity in comments would be impossible without it.

These data are relatively less suited to the second goal of reducing unintended bias regarding identity attributes. Namely, the data makes it difficult to test whether the ADS is biased against certain identities. The data does contain a number of features which describe the identity-related content of each comment. However, only a handful of these features have complete data. Most of the identity features are missing values for around 80% of the data. For the few features that are actually complete, the fraction of annotators that assigned each identity label to each comment is not included. Only a boolean vector representing whether the annotator fractions are greater than some unknown cutoff exist in the data.

Consequently, it becomes very difficult to assess potential bias in the data. Any metric calculated using identity features with many missing values will have a large degree of variance. Additionally, since there are very few identity features with complete data, there are few comparisons with which to actually assess fairness. For instance, the ADS may be able to return a low variance estimate of accuracy for homosexual-related comments. But if there are no other sexuality-related categories, it is difficult to assess whether homosexual terms are being disproportionately advantaged or disadvantaged.

5.2 Fairness of Implementation

Our analysis indicates the model is largely accurate, but unfair and biased for attributing protected features, such as race, to toxicity. We report AUC, Accuracy, False Positive Rates and False Negative Rates per subgroup to assess accuracy and treatment between subgroups. We analyze SHAP values to assess what features are influencing toxicity predictions.

As proposed previously, suppose this tool is used to moderate social media platform comments. The AUC scores and accuracy results are generally high, which is beneficial to platform owners who want to claim the tool works for all subgroups. The FPR's are low and FNR's are high, suggesting that the model is biased to predict non-toxic. This is beneficial to users who support less moderation and more freedom of speech. This could be harmful to users who are targeted for online harassment and want better moderation. The platform may or may not benefit from a lenient ADS depending on the general user-preference. FPR's and FNR's are similar across subgroups within each attribute, which would benefit a social media platform claiming fair treatment. However, the FPR rates for Black- and White-related comments are higher than other racial features, indicating a difference in treatment and possibly unfairness.

We further assess fairness by analyzing SHAP values for various comments, specifically when Black and White topics are mentioned. The explanations show that these features are not only attributed to toxicity, but they often have the most weight compared to other words. This demonstrates that although the model is accurate and "lenient", it is learning to be biased given the occurrence of a protected feature. This is an unreasonable and unfair association rule that can be harmful to unprivileged groups.

5.3 Potential Deployment

Since the ADS appears to associate mention of certain identities with toxicity, this ADS is not appropriate for deployment in any context. Deployed as-is, there is a high risk of those with particular identities being flagged as "toxic" simply for identifying with that group. This could serve to unfairly discourage or effectively disallow those belonging to a particular identity from commenting on platforms using the ADS.

5.4 Next Steps

Ultimately, the preprocessing by the data collectors, Jigsaw, causes many issues when assessing fairness. Many identity features have intentional missing values with no explanation. The identity features that don't have missing values are reduced to binary vectors rather than indicating the fraction of annotators that denoted the comment as pertaining to those identities. This affords little flexibility in understanding the necessarily subjective labeling by the annotators.

In addition to improving the completeness and transparency of the data, increasing the number of annotators may help reduce uncertainty in the annotated labels. Furthermore, deciding what counts as "toxic" is subject to biases in individual annotators–including more annotators may also serve to counteract the individual annotator bias.

The preprocessing done by the ADS is extremely limited. More text processing like removing stopwords, punctuation, and stemming may serve to reduce noise introduced by uninformative terms. Fully removing identity-related tokens may also be a way to avoid the ADS associating these terms with toxicity–however, this would require further investigation to understand if it's reasonable.

Finally, the ADS should introduce group fairness into its pipeline. Simply reporting a handful of subgroup AUC values is not sufficient. Including metrics like FPR and FNR can help identify disparate treatment across subgroups. Additionally, implementing some form of SHAP analysis can offer good spot checks to ensure identity-related terms are not being spuriously associated with toxicity.

The volume, velocity, and variety of toxic content generated today is alarming and should be a priority to many online platforms. Comment sections are a common place where hate speech and violence is explicitly exchanged between users. Therefore, the development of an ADS tool to moderate toxicity in comments is in demand; however, it would be counterproductive if such a tool only exacerbates harmful biases and further oppresses unprivileged groups.

References

[1] Borkan, Daniel. "Benchmark Kernel." Kaggle, March 28, 2019. https://www.kaggle.com/code/dborkan/benchmark-kernel/notebook.

[2] "Jigsaw Unintended Bias in Toxicity Classification." Kaggle, March 29, 2019. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/overview.

[3] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. Stanford, August 2014. https://nlp.stanford.edu/projects/glove/.