Analyzing Identity Mention Bias in Toxic Comment Classification

Katherine Hann[1] **Maddie Kalil**[1] **Sarah Payne**[1,2]
{khann22, mkalil, paynesa} @seas.upenn.edu

[1] Department of Computer and Information Science, University of Pennsylvania
[2] Department of Linguistics, University of Pennsylvania

Content Warning: This paper includes quoted comments from Wikipedia that contain racist, homophobic, sexist, and otherwise harmful language.

Abstract

Toxic comment classification is the task of determining the level of toxicity of comments on online platforms. Models capable of performing this task are in increasing demand due to the rise of abusive language and hate speech on various social media platforms. However, recent work has shown patterns of bias in these models; one such pattern is based on identities mentioned in the comments, with marginalized identity mentions increasing the likelihood of a toxic classification. Borkan et al. (2019) introduced five metrics for evaluating identity mention bias in toxic comment classification that extend the common evaluation metrics for classification. In this paper, we expand on these metrics by replicating Borkan et al. (2019) on a RoBERTa-based model for Wikipedia comment classification. We additionally introduce two novel metrics, stereotype bias and fuzzing, that provide complementary information to the existing five metrics, allowing for deeper analysis of identity mention bias in toxic comment classification.

1 Introduction

Due to the overwhelming volume of users, online platforms like Twitter now employ toxic comment classification to automatically determine a post's level of toxicity, so that exceedingly toxic comments can be flagged or removed. A comment may be toxic due to identity-based hate speech, or generally rude or offensive language. Having accurate automatic tools for detecting such comments on platforms like social media can help hold users accountable for their actions, as well as prevent future incidents from the same user; a highly accurate system would presumably also dissuade people from posting such content to begin with. In recent years, there has been increasing pressure on social media companies to provide this automatic classification in order to prevent cyberbullying and help protect users of all identities. Toxic comment classifiers have proven to be highly accurate under certain circumstances (e.g. Georgakopoulos et al. 2018; Schmidt and Wiegand 2017). However, most of them are trained on Standard American English (SAE), which is far from representative of the wide range of dialects present in the United States alone (Sap et al., 2019; Blodgett et al., 2016). Further, the training data itself may contain implicit biases (Davidson et al., 2019; Sap et al., 2019), potentially as a result of the biases of the human annotators. This leads to disparities in the performance of the models for certain demographics: if data or models are sensitive to dialect, comments in African American Vernacular English (AAVE) are more likely to be classified as toxic (Sap et al., 2019; Davidson et al., 2019; Blodgett et al., 2016). Meanwhile, if they are sensitive to mentions of marginalized identities, who tend to experience higher levels of cyberbullying, they may classify nontoxic comments mentioning these identities as toxic (Borkan et al., 2019; Dixon et al., 2018).

Recent work has made strides in evaluating the biases in toxic comment classifiers, especially with regards to Standard American English (SAE) vs. AAVE comments (e.g. Blodgett et al. 2016; Davidson et al. 2019; Sap et al. 2019). Identity mention bias, or the bias to classify comments mentioning marginalized identities as more toxic, is still a relatively new field of research, however. Biases such as these might cause, for example, "I identify as a queer Latina" to be classified as more toxic than "I identify as a white man" since the former comment implicates several marginalized identities (queer,

Latinx, female) and the latter does not. Borkan et al. (2019) propose five metrics that expand on standard evaluation for toxic comment classification models in order to better analyze these biases. In this paper, we replicate each of these five metrics to analyze a RoBERTa-based model for Wikipedia toxic comment classification. Additionally, we introduce two new metrics (stereotype bias and data fuzzing) that build on, and provide complementary information to, the metrics of Borkan et al. (2019).

2 Related Work

The state of the art for toxic comment classification models involves applying both SVM and RNN architectures combined with character-level n-grams. However, many challenges remain in this domain: there is no current benchmark dataset for hate speech detection, and because much of the research is extremely Anglocentric, focusing almost exclusively on English data, the current state of the art may not generalize well to other languages (Schmidt and Wiegand, 2017). As with many machine learning models, another challenge with comment classification models is bias. Previous work incorporates several different methods of evaluating bias in toxic comment classification models. Example metrics include Error Rate Equality Difference, which is a definition of fairness that is satisfied when the false positive rates and false negative rates are equal across comments containing different identity terms; Area Under the Receiver Operating Characteristic Curve (ROC-AUC), which provides an aggregate measure of performance across all possible classification thresholds; and Average Equality Gap, which is a graphical measure of the difference between the true positive (negative) rate of a subgroup and the true positive (negative) rate of the background. Several modifications have been proposed to ROC-AUC; these variations include Background Positive Subgroup Negative and Background Negative Supgroup Positive AUC (Borkan et al., 2019), as well as Pinned AUC (Dixon et al., 2018). Dixon et al. (2018) also propose the use of Pinned AUC Equality Difference, a measure of how similar pinned AUC scores are across different subgroups. More similar pinned AUC values mean similar performance within the overall distribution, indicating a lack of unintended bias. Each of these metrics have their own strengths and weaknesses and are used to analyze a range of common classifier biases, such as low subgroup separability and score shifts (Borkan et al., 2019).

Other research focuses on identifying potential causes of bias in these models. Sap et al. (2019) describe how data that is used to train automatic hate speech detection models is racially biased because of annotators' insensitivity to differences in dialect. They examine the example of African American English, finding that tweets with this dialect were up to two times more likely to be labelled as offensive compared to others; thus, models trained on these corpora acquire and propagate the biases of the annotators. Sap et al. (2019) provide similar findings on Twitter data for which they use Blodgett et al. (2016)'s model of demographic prediction based on geotagging of Tweets, which uses a separate language model for each dialect on which it is trained and predicts the most probable dialect for a given Tweet. Sap et al. (2019) additionally consider a data set in which race was self-reported by about 5,000 Twitter users, and find similarly strong correlations between demographic information and predicted toxicity. Finally, Sap et al. (2019) conducted a study through Mechanical Turk in which they had annotators label Tweets, but they were given the dialect or racial background of the Tweet's author. In this context, the annotators were less likely to label AAVE speech as offensive, and more likely to rate tweets as offensive to someone, but not offensive to themselves.

Another subset of research explores different strategies for mitigating bias in abusive language detection models. One successful strategy proposed by Park et al. (2018) involves using debiased word embeddings, which are pre-trained word embeddings that apply different techniques to remove bias associated with certain words, and gender swap data augmentation, in which male and female identities in the training data are swapped, thereby removing the correlation between gender and the classification decision. These modifications yielded a significant decrease in the False Positive Equality Difference accompanied by an only moderate decrease in the model's overall performance. Park et al. (2018) also experimented with fine-tuning the model with a larger, less-biased corpus, employing transfer learning to reduce the model's bias, which also proved effective in decreasing the False Positive Equality Difference score. In the present work, we will focus on evaluation methods for toxic comment classification bias, but it is our

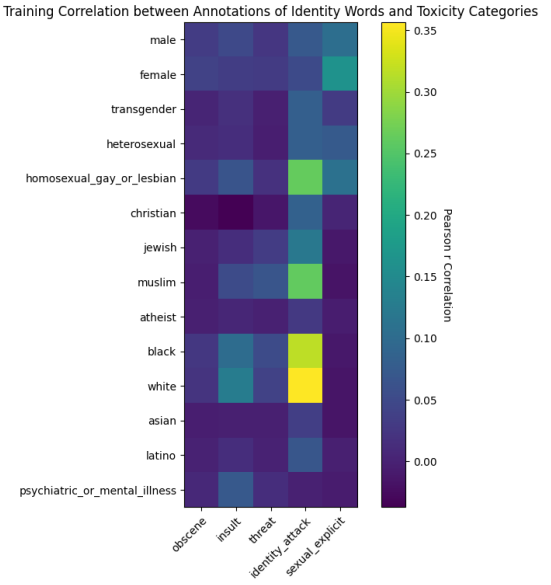

Figure 1: Pearson's r correlations between identity annotations and toxicity categories in the training data

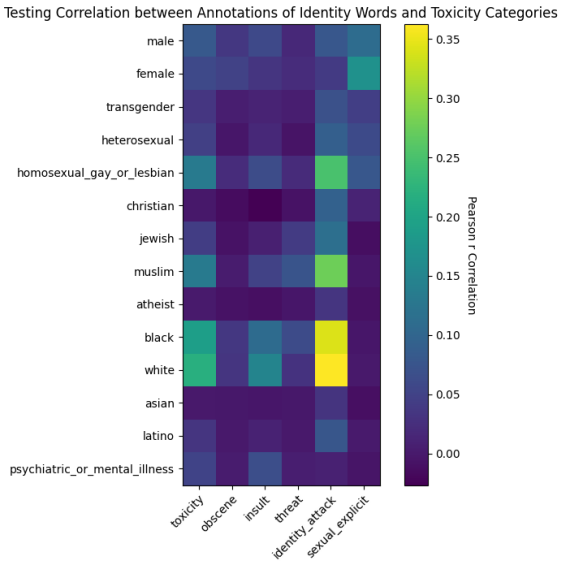

Figure 2: Pearson's r correlations between identity annotations and toxicity categories in the testing data

hope that a more robust suite of evaluation methods will spur future work on debiasing these models.

3 Design

3.1 Data

We take our data from the Kaggle Unintended Bias in Toxic Comment Classification Challenge,[1] which contains annotated comments from Wikipedia's talk page edits. Annotators marked the comments as mentioning or not mentioning a number of different identities (see Figure 1 for a full list) and for the toxicity categories `severe_toxicity`, `obscene`, `threat`, `insult`, `identity_attack`, and `sexual_explicit`. For the current work, we only consider identities for which there are 100 or more positive examples of the identity in our testing data, so that the results of our analysis will not be overly sensitive to only a few examples. The identities we therefore exclude from our analysis are: `other_gender`, `bisexual`, `other_sexual_orientation`, `hindu`, `buddhist`, `other_religion`, `other_race_or_ethnicity`, `physical_disability`, `intellectual_or_learning_disability`, `other_disability`. The lack of `hindu` and `buddhist` mentions in this data can be explained by their relatively smaller numbers of adherents in the Western world, from which these comments are mainly drawn since they are in English, meaning that they have less media presence. The remaining identities' exclusion can be explained by their lack of easy identifiers in a comment. In fuzzing the data for section 4.4, this hypothesis was confirmed, because even after performing the filtering to find the most common words for an identity, there were no unifying characteristics of any of these identities' top three most common words. So overall, removing these identities from the analysis is useful. We additionally exclude `severe_toxicity` from our analysis since this category contains too few positive examples for most evaluation measures. The training data contains 1,804,874 comments, 235,087 of which have human annotations for toxicity and identity mentions. For our testing data, we combine the public and private test sets provided by Kaggle and only consider those comments for which there are human annotations, giving a test set of 24,629 comments. Following Sap et al. (2019), we compute the correlation between annotations of each identity group and each toxicity class using Pearson's r correlation, with the results in Figure 1. We additionally compute these correlations for our testing data, given in Figure 2.

As can be seen in Figures 1 and 2, there are

[1] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

Positive PMI of Training Data

	toxicity	obscene	insult	threat	identity_attack	sexual_explicit
male	0.069	0.388	0.22	0.234	0	0.618
female	0.004	0.583	0.252	0.177	0	0.941
transgender	0	0	0	0	0.183	0
other_gender	0	0	0.798	0	0	0
heterosexual	0	0	0	0	0.263	0.886
homosexual_gay_or_lesbian	0	0	0	0	0.309	0.229
bisexual	0	0	0	0	0.138	0.578
other_sexual_orientation	0	0	0	0	0.278	0
christian	0.689	0	0.05	0	0	0
jewish	0	0	0	0.545	0.237	0
muslim	1.023	0	0	0.326	0.3	0
hindu	0	0	0.025	0.104	0.08	0
buddhist	0	0	0	0	0.269	0
atheist	0	0.263	0.142	0	0	0
other_religion	0	0	0	0	0.26	0
black	0	0	0	0	0.308	0
white	0	0	0	0	0.187	0
asian	0	0.358	0.119	0.005	0	0
latino	0	0	0	0	0.25	0
other_race_or_ethnicity	0	0	0	0	0.364	0
physical_disability	0	0	0.511	0	0	0
intellectual_or_learning_disability	0	0	0.798	0	0	0
psychiatric_or_mental_illness	0	0	0.649	0	0	0
other_disability	0	0	0.798	0	0	0

Figure 3: Positive PMI, measuring the association between identity mentions and toxicity labels in the training data

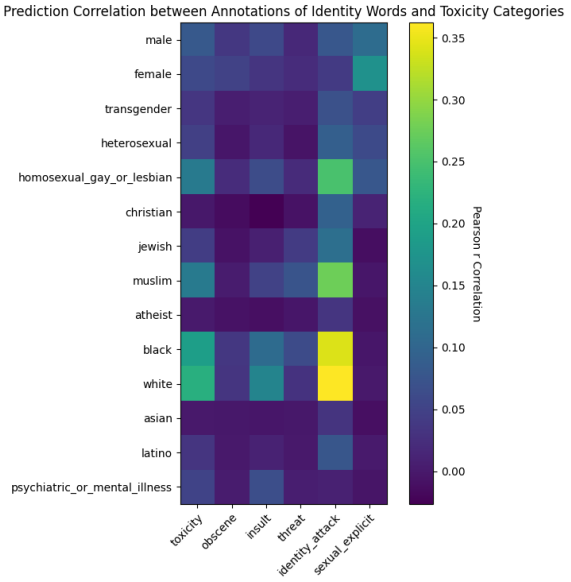

Figure 4: Pearson's r correlations between identity annotations and predicted toxicity categories

some relatively strong correlations between some marginalized identities and toxicity tags in both the training and testing data. For example, there is a correlation between `identity_attack` and `black` as well as `identity_attack` and `muslim`, and to a lesser extent, `jewish`. `identity_attack` is also correlated with `homosexual_gay_or_lesbian`. Furthermore, `female` is correlated with `sexual_explicit` more than `male`, which is unsurprising. Interestingly, `white` is also correlated strongly with `identity_attack`, which does not seem as logical; reasons for this correlation will be discussed further in section 4.1.2.

We additionally calculate Positive Pointwise Mutual Information (PPMI) over the training data. The PPMI calculations from Figure 3 indicate which identity mention - toxicity label pairs co-occurred more frequently than we would expect if we instead consider the identity mentions and toxicity labels as independently occurring events in the training data. Positive values correspond to pairs that co-occur more frequently than our expectation, while zero values correspond to pairs that co-occur less frequently. The highest PPMI values occur with `muslim` identity mentions and `toxicity` labels, and with `female` identity mentions and `sexual_explicit` labels. The PPMI values offer further insight into the correlations between toxicity labels and identity mentions present in the training data, which may also help explain any unintended bias in the model, since the model may learn to associate a certain identity mention with a certain toxicity label if the amount of co-occurring examples in the training data is disproportionate.

In analyzing the training data, it would also be interesting to consider how the identities of the annotators themselves factored into their annotations, but unfortunately, such information about the annotators is not provided in the data set.

3.2 Model and Preliminary Evaluation

We perform our analysis on the pre-trained "unbiased" Detoxify model from HuggingFace,[2] which was fine-tuned on the Kaggle training data shown in Figure 1. Pre-training of the model is done with RoBERTa (Liu et al., 2019), and it is fine-tuned with an Adam optimizer using a learning rate of 3e-5, weight decay of 3e-6, and a maximum of 100 epochs. The model utilizes multi-headed self-attention with 12 heads.

For our initial analysis, we consider the correlations between the model's predictions and the annotated identities in each category, using Pearson's r correlation as we did for the training and testing data; results are in Figure 4. As can be seen, the correlations for the model's predictions retain most of the high correlations in the training and testing data discussed above.

[2]https://huggingface.co/unitary/toxic-bert

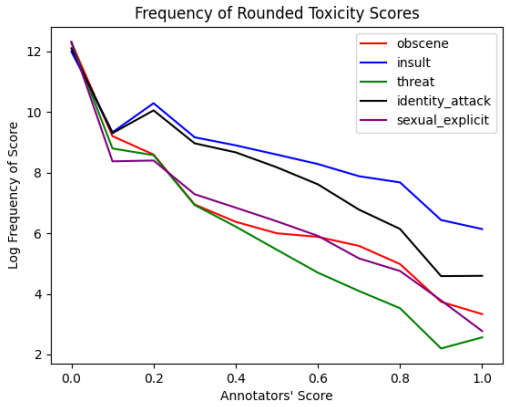

Figure 5: Log frequencies of each annotated toxicity score (rounded to tenths) for each toxicity label

As a second preliminary step in evaluation, we consider the ROC-AUC measure for each toxicity category. This measure calculates the area under the ROC curve, which is a measure of True Positive Rate (TPR) vs. False Positive Rate (FPR), where:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

Higher values of ROC-AUC correspond to greater area under the curve and better discrimination by the model; an AUC of 1 corresponds to perfect classification while one of 0 corresponds to failing entirely to distinguish between classes (Narkhede, 2018). ROC-AUC requires that the gold-standard values be binarized, and we binarize the values in the Kaggle data by taking a threshold of 0.5, as suggested by the authors of the challenge. To verify that this is a suitable threshold, we examine the frequencies of each annotated score, rounded to tenths, in Figure 5. As can be seen in this figure, there is a fairly steadily decreasing frequency for each score after 0.1, giving us no clear threshold from this analysis. There are local minima at 0.1 for several of the toxicity classes, making this appear to be a plausible threshold. However, of the 24,629 comments, 9,422 have annotated toxicity above the 0.1 threshold. This means that 38.25% of the comments would be classified as toxic. Based on qualitative understanding of online commentary, this is far too high. Furthermore, we find throughout our analysis that a threshold of 0.5 for binarization leads to better model performance on analysis tasks than 0.1, and as such, we use 0.5 as our threshold as suggested by the authors.

Toxicity Type	ROC-AUC Score
`toxicity`	0.94900
`obscene`	0.99474
`threat`	0.97261
`insult`	0.97092
`identity_attack`	0.96248
`sexual_explicit`	0.98904

Table 1: Overall ROC-AUC scores for each toxicity class (except `severe_toxicity`, for which there were insufficient positive examples)

Table 1 gives the ROC-AUC scores for each of the toxicity classes we consider. We can see from Table 1 that by the general AUC measure, the model performs well on all toxicity classifications, achieving at least .94 AUC and succeeding particularly well on `obscene` and `sexual_explicit`. However, these general AUC measures cannot give us information about unintended identity biases in the model, and in the next section, we turn to other measures which give us more insight into the biases of the model.

4 Results

4.1 Modified AUC Measures

We use the three modified ROC-AUC measures given by Borkan et al. (2019): subgroup AUC, background positive subgroup negative (BPSN) AUC, and background negative subgroup positive (BNSP) AUC. For each of these evaluation metrics, "subgroup" refers to the identity group currently under consideration, and "background" refers to all other groups. Subgroup AUC calculates AUC on only examples from a single identity, while BPSN calculates AUC over the positive examples in the background and negative examples in the subgroup and BNSP calculates AUC over the negative examples in the background and positive examples in the subgroup. More concretely, let S^+ be the positive examples in the subgroup and S^- be the negative examples in the subgroup, and let B^+ be the positive examples in the background and B^- be the negative examples in the background. Then:

$$subgroup.AUC = AUC(S^+ + S^-)$$

$$BPSN.AUC = AUC(B^+ + S^-)$$

$$BNSP.AUC = AUC(B^- + S^+)$$

We discuss each of these measurements and their results in turn below.

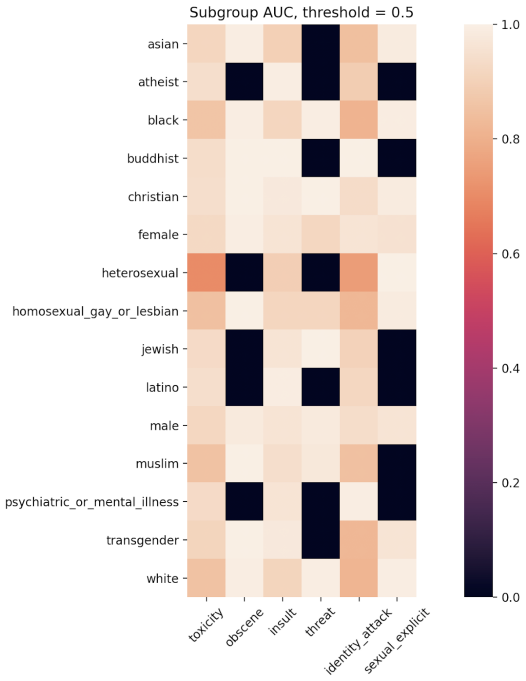

Figure 6: Subgroup AUC values for each identity-toxicity pairing

4.1.1 Subgroup AUC

Subgroup AUC calculates AUC on only examples from a single identity, which is a good representation of how well the model separates positive and negative examples within a subgroup. As discussed by Borkan et al. (2019), subgroup AUC score helps capture the type of bias called low subgroup separability, which is when a classifier underperforms on a subgroup relative to the background distribution. We found that a significant portion of the subgroup AUC values could not be computed (those shown as black squares) because those particular sets of comments had only one class present in the true labels – either all comments were annotated with toxicity scores above 0.5 (positive examples), or all comments were annotated with toxicity scores below 0.5 (negative examples). In fact, each of the non-computed values came from a subgroup where all the comments in which that identity was mentioned were annotated as non-toxic. The overall percentage of toxic comments in the test set using a threshold of 0.5 is only 5%, so when we further divide the data by identity mentions it would make sense that there are groups of comments which contain all very low toxicity scores.

Overall, the average subgroup AUC score across all identities is very close to 1. This indicates that the model performs well at separating toxic and non-toxic comments within each identity, and also matches the subgroup AUC results that Borkan et al. (2019) found in their analysis of the Perspective API toxicity models. The lowest subgroup AUC scores were for classifying regular toxicity in comments that mentioned `heterosexual` and `homosexual_gay_or_lesbian` identities, but with values in the range of 0.7-0.8, these are still decent subgroup AUC scores.

However, the AUC score of the model on a strictly per-group identity dataset perhaps does not effectively identify unintended bias in the model. One reason for this is that the AUC is classification-threshold-invariant, meaning that it "measures the quality of the model's predictions irrespective of what classification threshold is chosen".[3] In cases where the cost of false positives vs. false negatives is very different, we may want the model to focus on minimizing one type of error over the other. In the case of toxic comment classification, the trade-off is between incorrectly classifying benign comments as toxic, versus letting some toxic comments escape detection. Subgroup AUC doesn't take into account this tradeoff, and thus observing subgroup AUC in conjunction with other scores discussed in the following sections gives better insight about the model's unintended bias.

4.1.2 Background Positive Subgroup Negative AUC

Background positive subgroup negative (BPSN) AUC measures AUC on positive examples (annotated as toxic) in the background, and negative examples (annotated as non-toxic) in the subgroup. Borkan et al. (2019) note that BPSN AUC is particularly sensitive to false positives, since it considers only negative examples from the subgroup and a large number of false positives affects the number of true negatives in the model's predictions. BPSN AUC scores for all identity groups and toxicity classes are given in Figure 7. Of all BPSN AUC values computed, the three highest are for `obscene atheist` (0.99836), `obscene transgender` (0.99660), and `obscene asian` (0.99635), which seems to indicate that the model does well overall with classification of obscenity, perhaps because there are more salient lexical cues to obscenity than the other toxicity categories. There are 9,603 obscene comments in the training data, and manual inspec-

[3] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

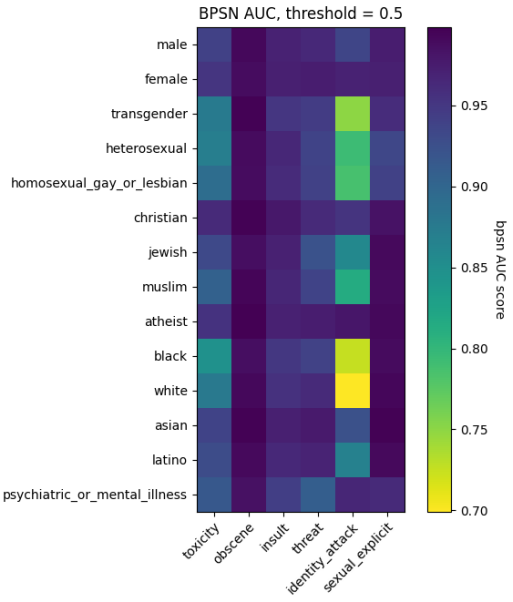

Figure 7: Background Positive Subgroup Negative AUC values for each identity-toxicity pairing

tion finds that many of them indeed include such lexical cues:

> "Like I said, you don't know $het dikwad!"
>
> "You are an ungrateful, selfish son of a b*tch."

Both of these quotes include lexical cues (in the form of profanity, although the former one uses alternate spellings) that would aid significantly in their classification as being obscene.

Of all BPSN AUC values computed, the three lowest are: `identity_attack white` (0.69930), `identity_attack black` (0.72616), and `identity_attack transgender` (0.74966), indicating a pattern of the model performing poorly on `identity_attack`; we can see in Figure 7 that BPSN is relatively quite low in this column. The identities for which `identity_attack` BPSN is lowest are mainly marginalized: `transgender homosexual_gay_or_lesbian, jewish, muslim, black` and `latino`. However, the lowest of all is for `identity_attack white`, and BPSN AUC is also low for `identity_attack heterosexual`. It makes sense that the model would yield more false positives for identity attacks on marginalized identities given that they are more prone to identity attacks online, but these latter two seem illogical at first glance.

Looking more closely at the training data, we find that 25,082 comments are annotated as `white`, and of these 3962 are annotated as `identity_attack`. Further inspection finds that many of these comments are related to "reverse racism," or the belief that white people are unfairly blamed for the current problems in the United States and world. Such feelings often manifest in highly racist, homophobic, or misogynistic comments, for example:

> "Black people seem to be waiting to find something wrong with white people to perpetuate their myth that all their problems are white people."
>
> "We know all women are liars and men (at least white christian ones) don't beat their families."

In our inspection of the comments, we did not find any that expressed truly anti-white views, but rather all comments annotated as `identity_attack white` expressed the sentiments discussed above, often at the cost of those of marginalized identities such as Black folks and women as seen in the examples. A correlation also existed in our training and testing data between annotations of whiteness and `identity_attack` (Figure 1), which now makes much more sense given our manual inspection of the data.

Closer inspection of the training data also reveals that there are 1,291 comments annotated as `heterosexual` of which 185 are annotated as `identity_attack`. Taking a closer look at these comments, we find that a large number of them express homophobic views by making comparisons between straight and queer folks, for example:

> "Heterosexuals do not have sex with juvenile males. Gay pedophiles do."
>
> "Homosexual acts aren't ordered to the natural functions of the body."

The low BPSN AUC for `heterosexual identity_attack` thus likely results from the fact that many of the examples annotated for these in the training data express homophobic views while also mentioning heterosexuality. The false positives leading to the low BPSN AUC thus likely result from the model overapplying this pattern of

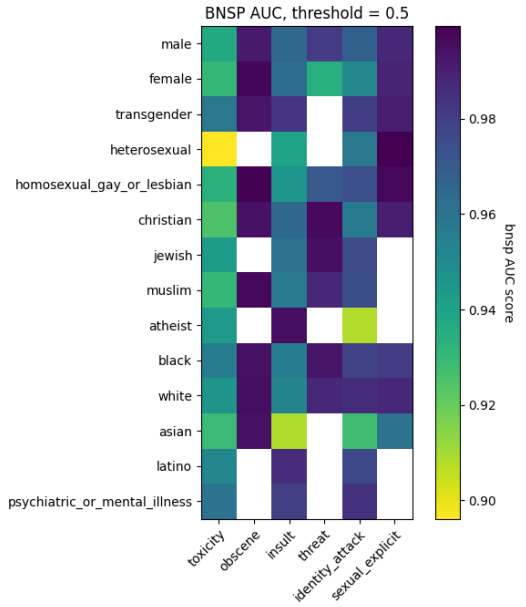

Figure 8: Background Negative Subgroup Positive AUC values for each identity-toxicity pairing

heterosexual-mentioning posts attacking homosexual identities, and thus falsely classifying some nontoxic posts in the test set that mention heterosexuality as toxic. In sum, BPSN AUC is particularly low for marginalized identities for `identity_attack` since the model overgeneralizes the patterns of these identities being more likely to be attacked; for non-marginalized identities that have low BPSN AUC, this results from identity attacks towards other marginalized identities that occur in comments also mentioning these identities.

4.1.3 Background Negative Subgroup Positive AUC

Background negative, subgroup positive (BNSP) AUC measures AUC over the positive (toxic) examples in the subgroup and the negative (non-toxic) examples in the background. As Borkan et al. (2019) note, BNSP AUC is related to the number of false negatives the model predicts for the subgroup, since these will be directly removed from the subgroup positive group. As can be seen in Figure 8, there are many identities and toxicity types for which AUC is not computable; this is caused by cases where there are insufficient negative or positive examples caused by either the background or subgroup being too small. Such cases are represented as white squares in Figure 8. Note that the range of BNSP from 0.89 to 0.99 is far better than that for BPSN, indicating that the model has more of a problem with false positives than false negatives. Of all BNSP AUCs calculated, the model performs the best on `sexual_explicit white` (0.98770), `sexual_explicit homosexual_gay_or_lesbian` (0.99746) and `sexual_explicit heterosexual` (0.99950), indicating that the model produces few false negatives with this category. By contrast, the model performs worst on `toxicity heterosexual` (0.89606), `insult asian` (0.90827), and `toxicity christian` (0.92501).

Compared to the extremely low BPSN scores for `white` and `black` above, `asian` fared relatively well for BPSN AUC, indicating that there were not a high number of false positives – in fact, it had the best BPSN AUC score of any of the races. However, here the BNSP score for `asian` is quite low across toxicity categories (save `obscene`, which is met with success again likely due to the lexical cues discussed above) meaning that errors that the model makes with this identity group are fairly consistently errors of false negatives. Manual inspection of the training data finds that there are 4578 comments annotated as `asian`, 291 of which are annotated as insults. These comments vary greatly in content: some focus on politics (e.g. "When Tillerson was in China and the Chinese had agreed to impose sanctions on North Korea, everything appeared to be headed a positive direction") and are typically insulting towards those who disagree with the political views being put forth rather than folks of a certain identity. However, there are also severely racist comments present:

> "Somebody better tell the asians and indians that they are too dumb to grasp all those heady math concepts!"
>
> "Asians do not whine. Most browns do not whine. Just blacks"

In which the former comment expresses anti-Asian racism and the latter expresses anti-Black racism, perpetuating the "model minority" stereotype.

It is possible that the reason that the model suffers from false negatives on data annotated as having Asian identity mentions is that the comments are more varied in scope and topic than those of other identities discussed above. It is also possible

that annotation bias may come into play here: it is interesting that Asians are the only racial group for which there is such a relatively low BNSP AUC and that they are also a racial group that is typically viewed as a "model minority." This has come to particular attention in the past year as more people have become aware of anti-Asian discrimination in light of Covid-19. It would be interesting – and likely very sobering – to run a similar analysis to the present one on more recent data to see how annotations and biases in the data change as a result of the surge in anti-Asian racism.

Another area where BNSP is particularly low is overall `toxicity heterosexual`. This is interesting in light of the low BPSN for heterosexual identity attacks discussed above, indicating false positives there. It is thus apparent that the model overgeneralizes the pattern of heterosexuality occurring with (typically homophobic) identity attacks, while simultaneously predicting false negatives for this non-marginalized identity. Inspection of the training data finds that of the 313 comments annotated as mentioning heterosexuality and being some type of toxic, over half (185) are annotated as `identity_attack`, with the second-most frequent annotation being insult (86), followed by `sexual_explicit` (37), `obscene` (4), and `threat` (1). It thus seems that the model overgeneralizes for the most frequent toxic label occurring with `heterosexual` but not for the rest, resulting in the low overall BNSP score observed here. Manual inspection of the comments tagged as `heterosexual` but not identity_attack finds that they are more related to the Church's handling of homosexuality and the supreme court case leading to the legalization of gay marriage in the United States; for example:

> "We made up the natural order and wrote gays out. God did not."

The latter comment does not represent a homophobic view but is labeled `insult` (presumably to Christians or Republicans later in the comment). It thus seems that the model overgeneralizes `heterosexual` when the label is a reliable indicator of homophobic attacks (i.e., for `identity_attack`), but not elsewhere. Thus, despite the model having overall very high BNSP AUC scores, it still struggles with false negatives both for Asian-mentioning comments – presumably due to their scattered subject matter or because of the view of Asians as a "model minority" – and with non-marginalized identities such as `heterosexual` and `christian`.

4.2 Modified AEG Measures

The Average Equality Gap (AEG) is the difference between correct classification rates of the subgroup and correct classification rates of the background, at a specific threshold. For Positive AEG, we are therefore comparing the true positive rates (TPR) of the subgroup to the background, and for Negative AEG, we are comparing the true negative rates (TNR) instead. The value computed by AEG ranges from -0.5 to 0.5, where -0.5 represents bias where the true rate of the subgroup is consistently higher than that of the background, and 0.5 represents bias where the true rate of the subgroup is consistently lower than that of the background. A value of 0 is best since it means the subgroup and background distributions have identical means. For AEG, we have the idea of Equality of Opportunity, which is "the idea that individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome."[4] Here, we say that if Equality of Opportunity held at every threshold (0-1) for the classifier, then AEG = 0.

4.2.1 Positive AEG

Positive AEG in particular is the metric comparing positive examples from the subgroup with positive examples from the background with the goal of low separability, (the opposite of AUC in which we had the goal of few mis-orderings or high separability). A lower value Positive AEG means higher TPR in the subgroup than background, and a higher value AEG means lower TPR in the subgroup than background. Note that most results, 73% in fact, shown in Figure 9 have a positive AEG value less than the absolute value of 0.15, meaning that most subgroups' TPRs have little bias when compared with their background (the subgroup and background distributions have similar means, so there is little model bias for particular identities). However, the values that are greater than that absolute value are mainly all negative. This means that there is higher TPR in the subgroup than background for a few examples. Since the maximum value for AEG is 0.5, we will look at examples where the absolute value of the AEG score is at least 0.25. This includes `transgender obscene`, `female threat`, and `asian sexual_explicit`,

[4] https://ai.googleblog.com/2016/10/equality-of-opportunity-in-machine.html

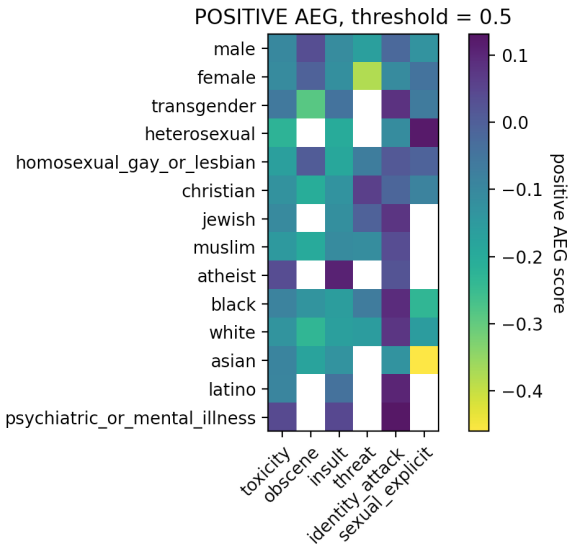

Figure 9: Positive AEG values for each identity-toxicity pairing

with `asian sexual_explicit` having the highest scores at -0.4604. Since having high negative Positive AEG scores indicates higher TPR in the subgroup than background, this actually agrees with BNSP in a roundabout way because we know that the `asian` identity has much lower scores for `sexual_explicit` than the background data. This is evidenced by the means: for all positive examples of `asian`, the average of the `sexual_explicit` scores is 0.00659, while for the data overall, the average for `sexual_explicit` is 0.01103. Thus, on average, the background is 1.5 times more sexually explicit than the Asian identity. So, it makes sense that the model would predict more true positives for `asian sexual_explicit` than the background because there are fewer true positives for this identity in general, so this does not necessarily represent a concerning model bias.

The more negative score for `female threat` may, on the other hand, represent bias, since it is not a result of the lack of toxic content for this identity that results in more subgroup TPRs. This higher TPR for `female threat` is likely due to a cultural and historical association of women with being threatened by men, which may explain why the average score for `female threat` is higher than the background `threat` score. Consider two examples from the data:

> "...im sure you can't wait to stone to death any women who speak up."

> "Explain why you still beat your wife and kids?"

There isn't the same association between silencing men or physically abusing husbands. These associations also carry throughout culture enough that they permeate different religions and races as well, so it is clear that other identities aren't singled out to receive the same level of threat as the `female` identity.

Positive AEG's score for the last identity and toxic class mentioned, `transgender obscene`, indicates a bias that is not picked up on by AUC. In fact, AUC indicates little bias for this class/identity, but according to Borkan et al. (2019), this is not contradictory. Rather, AEG is just the only metric out of the ones used that can pick up on "small score shifts." Borkan et al. (2019) states that "this type of bias occurs when a machine learning classifier outputs a consistently higher (or lower) score for a subgroup than for the overall data distribution...the shift is not to the extent as to confuse positive examples from the subgroup with negatives examples from the background distribution... AEG is the only metrics which pick up this subtle form of bias." Thus, we can explain this and other slight differences between AUC and AEG in this way. To dive deeper into the reason for this small score shift of `transgender obscene`, we can examine it from a cultural perspective similar to `female threat`. This is best exemplified by the comment:

> "...him taking her into the men's room and a transgendered person using the women's restroom are two totally different scenarios. A male who identifies as female isn't going to have his penis on display in the female bathroom."

This and other `transgender` comments may be labeled "obscene" in part due to the content that is typically discussed on online forums about the `transgender` identity. A discussion that mentions this identity is likely to contain content about what being transgender entails, which some annotators seemed to believe was inherently obscene, likely due to mentions of the specifics of related body parts, like in the comment above. Thus, the model is biased towards classifying more `transgender` identity comments as `obscene` due to the learned bias of annotations.

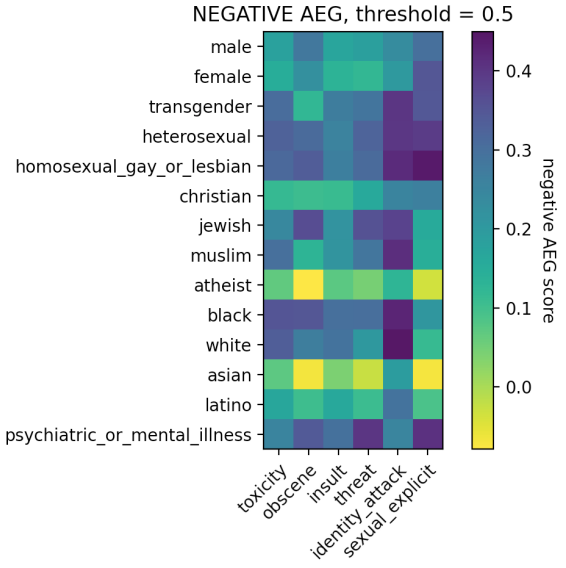

Figure 10: Negative AEG values for each identity-toxicity pairing

Our conclusions about the model's biases of `transgender obscene` and `female threat` may seem counter-intuitive based on the correlation chart. If this was a result of annotation bias, these identities/classes should appear to have higher correlation. However, we suspect that the bias still does result from the annotations, but is not reflected in the correlations graph because the model's bias comes from relatively few examples with really high scores. For example, for `female threat` we have that for all positive examples of female and threat, the average score is 0.66875.

4.2.2 Negative AEG

Negative AEG has the same goal of low separability as positive AEG, but now it compares negative examples from the subgroup with negative examples from the background. Negative AEG corresponds in some ways to BPSN AUC since the former considers true negative rates and the latter considers false positive rates. For any example for which the gold standard is negative, the model may either predict a true negative or a false positive; we can thus see negative AEG as providing a counterpoint to BPSN AUC by analyzing the half of the predictions for which negative is gold that BPSN AUC neglects. Thus, as expected, identities with low BPSN AUC (Figure 7) also have high negative AEG values (Figure 10). This is most apparent in the example of `white identity_attack` which has the highest negative AEG score and the lowest BPSN AUC score.

We first note that across the board, the absolute values of the negative AEG scores are higher than the absolute values of the positive AEG scores. For negative AEG, approximately half of the classes and identities have absolute values above 0.25. We also note that the scores are almost entirely > 0, meaning that the biased subgroups have lower TNRs than the background. The identities/classes for which BPSN AUC has a low score and Negative AEG has a high score indicate a shift towards toxicity of the classifier's scores for these identities, which are considered relatively non-toxic by the training data. For the identities on which BPSN AUC and Negative AEG do not have this agreement, (high BPSN AUCs and also high Negative AEG values) there is still an upward shift in the model's scores for relatively non-toxic items, but it is not large enough to cause mis-orderings with toxic items (Borkan et al., 2019). Thus, much of the bias displayed by the high Negative AEG scores shown in Figure 10 is actually relatively negligible, since it does not cause mis-orderings with toxic items. So we will focus on the four identities and classes that have the very highest Negative AEG scores, which are `homosexual_gay_or_lesbian sexual_explicit`, `white identity_attack`, `black identity_attack`, `transgender identity_attack`. Three out of these four agree with the results of BPSN AUC, and reasons for their bias are discussed in section 4.1.2, to which we defer for further explanation of these results. However, the bias of `homosexual_gay_or_lesbian sexual_explicit` is not reflected by BPSN AUC, so we will discuss this result here. There is little correlation indicated between `homosexual_gay_or_lesbian sexual_explicit`, however, we notice something about `homosexual_gay_or_lesbian` as a subgroup. The average number of identities mentioned in a comment (above the usual threshold of 0.5) is 2.007. For comments that are labeled as `homosexual_gay_or_lesbian` however, the average number of identities mentioned is 3.804. This is almost twice the number of identities in comments of this type, each of which has bias associated with it. So, on average, a comment that mentions `homosexual_gay_or_lesbian` has approximately three other identities mentioned

in the same comment, so we attribute this shift towards toxicity for this identity and class to the compounding effects of having a multitude of identities mentioned in a particular comment.

4.3 Stereotype Bias Measurement with WinoBias

The modified AUC and AEG measures introduced above give us a good sense of the false positives and negatives in our data and how these correspond to identity mention biases. However, the specific cues to these identities that the model is sensitive to are still largely a mystery. By considering pro- and anti- stereotypical sentences for differing identities and comparing toxicity ratings across them, we hope to gain understanding of how lexical cues and stereotypes play into the model's biases.

The WinoBias dataset was originally developed by Zhao et al. (2018) for detecting and evaluating gender bias in co-reference resolution models. The dataset consists of a number of sentences labeled as *pro-stereotypical* or *anti-stereotypical* based on labor department statistics for the jobs mentioned in the sentences. Each sentence is duplicated to use male and female pronouns so that for each pro-stereotypical sentence there is a corresponding anti-stereotypical sentence of the opposite gender formed by simply swapping a pronoun. The sentences in WinoBias fall into two categories broadly: Type 1 sentences are of the form **[entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances]**, while Type 2 are of the form **[entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances]**. As Zhao et al. (2018) note, Type 1 co-references cannot be resolved using syntax, while Type 2 may be resolved using only syntactic cues. For the case of toxicity classification, these cues are likely less important, however, and so we combine the test sets from each type to create one test set containing 792 (pro-stereotypical, anti-stereotypical) sentence pairings. These pairings are split evenly such that there are 396 pairings for which the anti-stereotypical sentence contains female pronouns and 396 for which it contains male pronouns. Because WinoBias was designed for a different task than the one at hand, these sentences are not intended to be toxic, and there are no gold-standard labels to which we may compare our model predictions. In the following, however, we wish to demonstrate that datasets such as WinoBias may still provide valuable insights into toxic comment classification performance.

Although WinoBias was not designed for toxic comment classification, were the model to consistently predict pro-stereotypical sentences to be more or less toxic than anti-stereotypical ones, this would be a sign of a problem in model performance and would give us valuable information regarding the lexical cues put to use by the model. In the case that the model consistently predicted pro-stereotypical sentences to be more toxic, this would signal that it has become hyper-sensitive to words associated with stereotypical employments of each gender in a similar way to its becoming hypersensitive to some identities in the above. If instead it consistently predicted anti-stereotypical sentences to be more toxic, then it may yield false positives for comments mentioning folks of marginalized identities that do not follow the stereotypes of such identities. In either case, we would have an indication of bias and an understanding of what the cues to this bias are. In the following analysis, we thus measure the rate at which the model predicts anti-stereotypical examples to be more toxic than pro-stereotypical examples.

Upon testing on all 792 pairings, we find that the HuggingFace model predicts anti-stereotypical examples to be more toxic almost exactly 50% of the time for each toxicity category (grey bars in Figure 11). This seems promising as it initially appears that the model is not showing either of the potential errors discussed above. However, since the examples are evenly split across gender, it may well be the case that the model accomplishes this apparent balance while still classifying all anti-stereotypical female sentences as more toxic than their pro-stereotypical male counterparts, as long as it compensates for this elsewhere. To check for such behaviour, we consider separately the 396 pairings for which the anti-stereotypical examples contain female pronouns and the 396 pairings for which the anti-stereotypical examples contain male pronouns. Here, we find that for most toxicity categories, the model classifies far from 50% of the anti-stereotypical examples as more toxic than their pro-stereotypical counterparts (Figure 11). For example, for the `insult` toxicity tag, we find that anti-stereotypical examples are classified as more toxic 91.2% of the time when they use female pronouns, vs. 9.2% of the time when they use male pronouns. `toxic`, `obscene`, `insult`, and

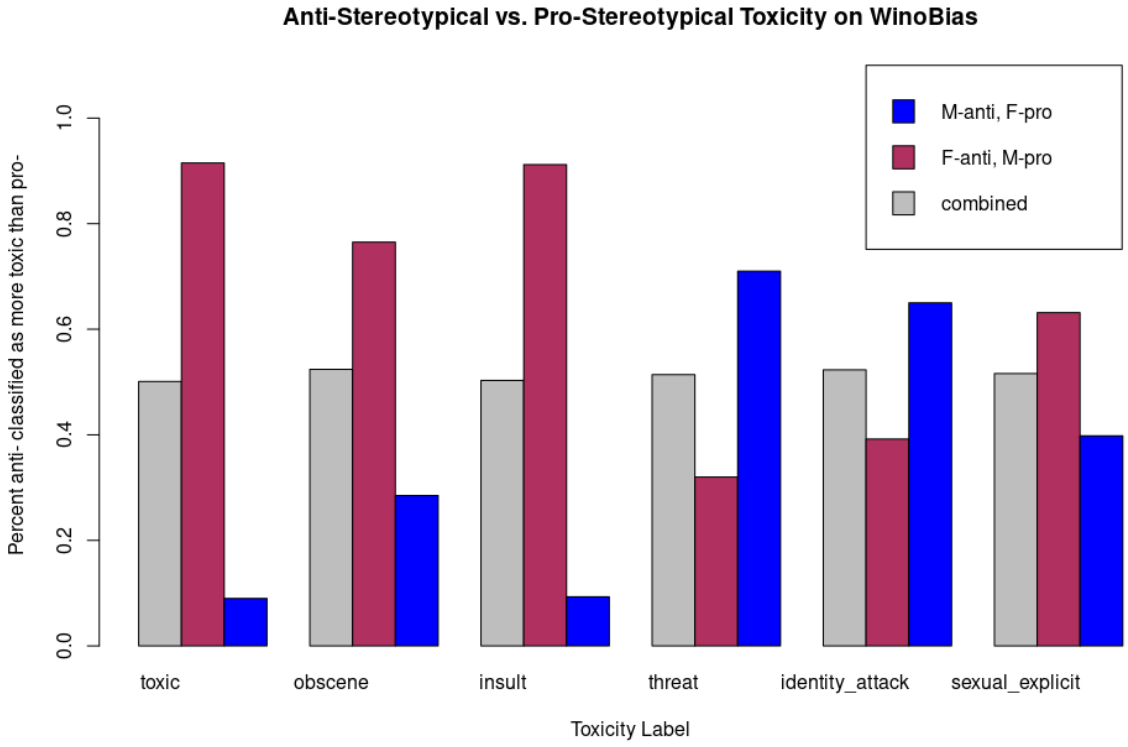

Figure 11: Percent of the time the anti-stereotypical examples are predicted as more toxic than the pro-stereotypical ones for each toxicity category

to a lesser extent, `sexual_explicit`, all have high rates of anti-stereotypical female-mentioning (henceforth F-anti) sentences being classified as more toxic than their M-pro counterparts, while M-anti are typically classified as less toxic than F-pro in these classes. In both cases, then, sentences mentioning female pronouns are classified as more toxic than the same sentences with the pronouns exchanged, showing that the model is particularly sensitive to gendered pronouns, regardless of whether the sentences they occur in are pro- or anti-stereotypical.

Interestingly, both `threat` and `identity_attack` show the opposite pattern: for these, male anti-stereotypical sentences are classified as more threatening and more likely to be identity attacks than the corresponding female pro-stereotypical sentences, and the female anti-stereotypical are classified as less toxic than the corresponding male pro-stereotypical sentences, meaning that the model is more likely to classify sentences with male pronouns as threats or identity attacks than those with female pronouns. This correlates with the slightly higher BPSN AUC found for `male identity_attack` above; the model predicting `identity_attack` more for `male` anti-stereotypical sentences may make sense in light of the fact that insults aimed at males often use feminizing, and thus anti-stereotypical, language.

Overall toxicity predication values of the model are fairly low for all of the WinoBias data (Table 2), but with relatively large standard deviation for most toxicity labels (Table 3). Unsurprisingly, the toxicity labels with lowest standard deviations are the ones for which the rates of anti-stereotypical examples being predicted as more toxic than pro-stereotypical examples are closest to 50% across all genders (Figure 11). We can thus see that looking beyond the overall performance on WinoBias gives us valuable information on biases in the HuggingFace model, despite the data being designed for a different evaluation task altogether. This analysis provides not only information regarding the lexical cues to which the model is sensitive, but also information regarding gender bias that is not as well-represented by the measures of Borkan et al. (2019).

	toxic	obscene	insult	threat	identity_attack	sexual_explicit
M-anti	0.0064	9.3e-05	0.0030	0.0015	3.8e-05	3.4e-05
M-pro	0.0079	0.0001	0.0054	0.0002	4.6e-05	3.6e-05
F-anti	0.0090	0.0001	0.0062	0.0002	4.9e-05	3.7e-05
F-pro	0.0067	9.5e-05	0.0032	0.0014	3.9e-05	3.4e-05
C-pro	0.0072	0.0001	0.0043	0.0008	4.2e-05	3.5e-05
C-anti	0.0076	0.0001	0.0045	0.0008	4.3e-05	3.6e-05

Table 2: Mean toxicity scores for anti- and pro- stereotypical WinoBias examples, for female-only (F), male-only (M), and combined (C)

	toxic	obscene	insult	threat	identity_attack	sexual_explicit
M-anti	0.0390	0.0003	0.0228	0.0194	5.7e-05	7.7e-05
M-pro	0.0468	0.0003	0.0401	0.0014	0.0001	6.5e-05
F-anti	0.0489	0.0003	0.0412	0.0013	0.0001	6.9e-05
F-pro	0.0408	0.0003	0.0248	0.0179	6.0e-05	7.8e-05
C-pro	0.0435	0.0003	0.0330	0.0126	9.5e-05	7.1e-05
C-anti	0.0439	0.0003	0.0330	0.0137	0.0001	7.3e-05

Table 3: Standard deviations for toxicity scores for anti- and pro- stereotypical WinoBias examples, for female-only (F), male-only (M), and combined (C)

Toxicity Type	Regular AUC	Fuzzed AUC
toxicity	0.94900	0.92918
obscene	0.99474	0.99489
threat	0.97261	0.96791
insult	0.97092	0.96712
identity_attack	0.96248	0.89466
sexual_explicit	0.98904	0.98717

Table 4: Regular and Fuzzed ROC-AUC scores for each toxicity category

4.4 Fuzzed Data

The fuzzed data set is yet another helpful tool for analyzing unintended bias in text classification models. By fuzzed data, we mean that we replace key words that identify identities in the data set with key words from other identities in the data set randomly. This allows us to gain a better sense of how much of the model's bias towards toxicity for certain comments is attributable to identity mentions in those comments. First we describe how we created the fuzzed data. The first step to fuzzing the data is determining the key words for each identity. We do this by tokenizing the comments and part-of-speech tagging them using NLTK. We then filter through the POS-tagged tokens, removing all punctuation and all stopwords. In completing this, we realized two patterns: almost all of the words that would be used by an annotator as a clue that a specific comment mentions a specific identity had the parts of speech of adjective ("JJ"), noun, ("NN"), or noun plural ("NNS"). Thus, we also choose to filter out all words that are not tagged with one of these parts of speech. The second pattern we noticed is that in the top 10 most common words for each identity, there were a few words that appeared across the board. These words were: "trump", "people", "many", "church", "god", and "jesus". The words trump, people, and many give no indication of what identities may be mentioned in that comment (we found that comments with "trump" were not always labeled with the male identity), so we chose to remove those. We removed "church" as it almost always appeared in a sentence containing "christian" or "catholic" and frequently appeared in the lists of identities not related to religion. We removed "god" and "jesus" because these have double meanings used for profanity, so they were not reliably related to the christian identity. We then took the three most common words from the lists of every identity and compiled them into two dictionaries, one mapping from the word to its part of speech one mapping from part of speech to the list of associated words. We then went through all of the comments and used regex to find occurrences of these words, and replace each one with a random word from the list of words whose part of speech matches the original. Part of the replacement code was drawn from ConversationAI's unintended ML bias analysis repository. [5] Some examples are as follows:

[5] https://github.com/conversationai/unintended-ml-bias-analysis/tree/main/unintended_ml_bias

Original comment 1: “Old white men ARE the swamp.”

Fuzzed comment 1: “Old black women ARE the swamp.”

Original comment 2: “say ‘Jews’ not ‘Jewish’”

Fuzzed comment 2: “say ‘muslims’ not ‘muslim’”

Original comment 3: “alt-white or all-white”

Fuzzed comment 3: “‘alt-chinese or all-chinese”

Note that this style of fuzzing does not work 100% of the time due to variation in the data. We still end up with comments that do not make total sense. For example, since “israel” is one of the identifier words for the `jewish` identity, and “china” is one of the identifier words for the `asian` identity, for a comment like “Miller is a sad man.” we ended up with a fuzzed comment of “Miller is a sad china.” To improve this in the future, we could add more specific find and replacement methods that account for people, places, and things, rather than just their part of speech. For our purposes, we believe it is okay if some of the comments don’t make as much logical sense since we want to determine how the words related to specific identities affect the toxicity of the model’s predictions.

If there is little or no unintended identity mention bias in the model, we hypothesize that there would be little or no difference in the subgroup AUC, BNSP/BPSN and AEG scores for the fuzzed and regular test sets. On the other hand, a more biased model may result in large differences in scores between the two test sets. The differences in scores could reflect the fact that the model is somehow using a specific identity term – which may be replaced during the fuzzing procedure – as a feature in determining the level of toxicity in a comment. Another indicator of bias is the model’s overall performance on the regular versus fuzzed test sets computed using AUC. Across all 6 types of toxicity, the AUC score decreased when we used the model to predict toxicity scores on the fuzzed data set, which seems to imply that the replaced identity word contributed significantly to the model’s predictions on the normal test set. The most drastic decrease occurs for `identity_attack`, which also fared poorly on BPSN AUC, indicating that the model may pay special attention to identity words here. Looking at the other metrics, we find that there are small differences between the fuzzed and normal test set across the scores for Subgroup AUC, BNSP/BPSN AUC, and AEG; in a majority of cases, the scores shifted in the direction that indicated worsening performance.

5 Discussion

In this paper, we have presented seven analysis methods, each of which provides complementary information for identity mention bias in toxic comment classification. Subgroup AUC helps us to see whether the model is underperforming within specific identity groups, and is the only measure from Borkan et al. (2019) that does not consider the background, but rather only considers separability within each subgroup. In contrast to this, the other four measures from Borkan et al. (2019) allow us to compare the performance of the model on a subgroup to that on the background, giving valuable information as to which identities the model fares particularly poorly with. While BPSN and BNSP AUC are particularly useful for analyzing patterns of false positives and false negatives, respectively, positive and negative AEG values provide comparative measures of true positive rates and true negative rates, respectively, between the subgroup and background. This can help us to deepen our understanding of whether the model’s underperformance on some categories as evidenced by BPSN AUC is evidence of an overall weakness of a specific bias for the given category. While complementary in their analysis, these five measures do little to explain the causes – lexical or otherwise – of model bias.

The used of a fuzzed dataset allows us to test certain lexical cues as being potential cues for bias in the model. For the model analyzed here, for example, we see via the Borkan et al. (2019) analyses that the model seems to be particularly underperforming on `identity_attack`, especially for marginalized identities. Our fuzzing analysis is able to verify that lexical cues strongly associated with certain identities are at least partially to blame for the model performance drop, since fuzzing these cues leads to a steep drop in performance relative to other toxicity categories. To further investigate specific lexical cues, the stereotype bias measurement presented in section 4.3 allows us to determine the model’s sensitivity to both gendered pronouns and to stereotypical roles for each gender, the former of which is much larger. In future work, it would be interesting to consider attention in the model to provide further insight into lexical cues that are heavily utilized by the model. Unfortu-

nately we were not able to investigate this in the current work due to the computational resources needed to fine-tune BERT-based models.

For the HuggingFace model considered in this paper, we have made several important observations: while the model does well with discriminating within each category, as evidenced by its relatively high subgroup AUC scores, it has a large rate of false positives for the `identity_attack` toxicity category, as evidenced by low BPSN AUC throughout this category and a large decrease in performance on the fuzzing dataset for this category. We found that this decrease in performance was not isolated to marginalized identity mentions, as white identity mentions often correlated with white supremacist comment content, which is clearly toxic, and heterosexual often correlated with homophobic content. This provides an important insight for future research in that the identities directly mentioned in comments may not always be the ones being attacked, and models may then become sensitive to these non-marginalized identities. We additionally find when considering the WinoBias dataset that the model is particularly sensitive to gendered pronouns, despite a strong difference in performance between male and female not being attested elsewhere. This suggests that relative performance on such swapped-word tasks as WinoBias has the capacity to reveal biases that metrics such as the Borkan et al. (2019) metrics cannot.

6 Conclusion

In this paper, we have tested a RoBERTa-based model for Wikipedia toxic comment classification against seven evaluation metrics, including modified ROC-AUC (Borkan et al., 2019), modified AEG (Borkan et al., 2019), data fuzzing, and stereotype bias (Zhao et al., 2018). We have demonstrated that each of these metrics provides complementary information regarding the biases of the model at hand: while the Borkan et al. (2019) metrics give us information about model performance on certain identity subgroups relative to the background, the latter two measures help give valuable information about the lexical cues that may contribute to such bias. One important takeaway from this work is that low model performance due to identity bias may occur for comments mentioning non-marginalized identities if these mentions are used to attack others (e.g. white supremacist comments). Another important conclusion is that lexical swapping such as that in the WinoBias data can reveal important clues to lexical biases (in this case, gendered pronouns) that may not be recognized by other error measures. It is our hope that these points and the seven evaluation metrics presented here will prove to be useful to future research on de-biasing toxic comment classification and other tasks throughout Natural Language Processing.

Acknowledgments

We are extremely grateful to Dr. Yatskar and Chaitanya for all of their helpful feedback and support throughout this project.

References

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

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

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

Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pages 1–6.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sarang Narkhede. 2018. Understanding auc-roc curve. *Towards Data Science*, 26:220–227.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

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.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.