Fair Model of Comment Toxicity

Arpita Singh
arpitasingh@umass.edu

Arundhati Gorkhe
agorkhe@umass.edu

Nandhinee Periyakaruppan
nperiyakarup@umass.edu

Arunima Sundar
asundar@umass.edu

1 Introduction

Reviewing user-generated text, eg. detecting toxic comments - is an important tool for moderating the text put up on the Internet. A fair classification of such text is very important, however, models can become biased due to learning various spurious features for all such comments. Our project deals with the CivilComments dataset, which has a set of comments and various attributes to it like toxicity and how severe it is. Prior work has shown that classification models do pick up on biases in the training data and spuriously associate toxicity with the mention of certain demographics.
Our aim is to apply fair learning methods to avoid such spurious assumptions made by the model. Our work will involve studying the performance of baseline classifiers, applying various fairness objectives to see which works best in this case, and study all of their outcomes to see if we can overcome the bias in a comment toxicity model.

2 Related work

(Google and Jigsaw, Perspective API, 2017) was created by Jigsaw and the Google Counter Abuse Technology team in a collaborative research initiative called Conversation-AI. The Conversation AI team worked on machine learning models that can identify toxicity in online conversations. PerspectiveAPI is built upon multiple public benchmark and real world datasets (including CivilComments) for different tasks (like multilingual toxic comment cla ssification and robustness evaluation). They found that these models incorrectly learned to associate the names of frequently attacked identities with toxicity. This happens because training data was pulled from real world scenarios where unfortunately, certain identities are overwhelmingly referred to in offensive ways. Training a model with this data led to spurious learning. This is an ongoing research, with the team trying to find ways to overcome this bias. In recent improvements with (Alyssa Lees, 2022), the new model outperformed the previous best performing model by approaching the problem with a single multilingual token-free Charformer model that is applicable across a range of languages, domains, and tasks. They reported better AUC-ROC scores and were able to extend the solution to 10 new languages.

In (Dixon et al., 2018), the authors introduce and illustrate a new approach to measuring and mitigating unintended bias in machine learning models. By using a set of common demographic identity terms as the subset of input features, the authors attempt to measure bias. This technique permits analysis in the common scenario where demographic information on authors and readers is unavailable, so that bias mitigation must focus on the content of the text itself. The mitigation method introduced is an unsupervised approach based on sub-sampling the training dataset. The authors demonstrate that this approach reduces the unintended bias without compromising overall model quality by measuring the model performance using various AUC score metrics.

The authors of (Reichert et al., 2020) describe a variety of toxic comment detection models while preserving high classification performance and reducing identity-driven bias. They trained the dataset using Logistic Regression, neural network and LSTM models whith different features like TF-IDF and GloVe embeddings, hyper-parameter configurations and different model architectures. Additionaly, they balanced the dataset using over-sampling and under-sampling techniques in combination with natural text generation. The authors demonstrated that rebalancing training

data helps in improving model classification.

In (Ji Ho Park, 2018), the authors measure gender biases on models trained with different abusive language datasets, while analyzing the effect of different pre-trained word embeddings and model architectures. The paper mentions various bias mitigation methods, and were able to effectively reduce gender bias by 90-98%. They believe the methods proposed in the paper can be extended to other identity problems as well.

3 Background and hypotheses

3.1 Background

1. B (Behavior) - correlation between the classification outcomes and mention of various protected groups

2. E (Environment) - the processes that produced the CivilComments dataset

3. S (System) - a comment toxicity classifier

4. T (Tasks) - accurate model that does not discriminate against the protected groups

3.2 Hypotheses and Research problem

After initial EDA, we ran a classical learning model (logistic regression) as our baseline on the given data, and observed that Comment Toxicity label is unfairly dependent on certain identities and protected attributes. The presence of certain terms in the comment could lead to a biased classification of the comment due to learning of spurious features by the model.

```
new_comment = "I'm gay"

# Do not change the code below
classify_string(new_comment)
```

```
TOXIC: I'm gay
```

Figure 1: Example of incorrect labeling due to spurious learning

It would be interesting to see which protected attributes contribute more to the bias which we plan on measuring using Demographic Parity Ratio.

Since we see a suspected discriminatory dependence of Z on Y, we think that the 'Marginal Interventional Mixture' fairness objective would be best suited for this problem. We will test this hypothesis by applying various fairness objectives and comparing their results.

4 Data

At the end of 2017 the Civil Comments platform shut down and chose to make ∼2M comments from their platform publicly available in a lasting open archive so that researchers could understand and improve civility in online conversations for years to come. Jigsaw sponsored this effort and extended annotation of this data by human raters for various toxic conversational attributes.The text of the individual comment is found in the comment_text column. The data contains the following datasets :-

1. train.csv - The dataset, which includes toxicity labels and subgroups. We've split this dataset further into train and test since there was enough data in the this csv to work with.

2. The training set contains 1804874 comments.

3. Table 1 defines all the identity groups present in the dataset and the number of times they appear cumulatively.

Identity Groups	No. of appearances in dataset
asian	1789
atheist	536
bisexual	817
black	7052
buddhist	258
christian	11482
female	14428
heterosexual	1224
hindu	323
homosexual_gay_or_lesbian	5807
intellectual_or_learning_disability	759
jewish	2927
latino	1603
male	16083
muslim	7490
other_disability	688
other_gender	644
other_race_or_ethnicity	3786
other_religion	3484
other_sexual_orientation	1191
physical_disability	630
psychiatric_or_mental_illness	2747
transgender	1580
white	10419

Table 1: No. of appearances of each identity in the comments dataset.

5 Techniques and Methods

Since we were dealing with unclean text data, we needed to use different preprocessing techniques to prepare the data before fitting the models. We removed all punctuation, numbers, emojis and stop words (using gensim) from the data. We also changed all text data to lower case, cleaned all contractions, and removed all special characters. We also performed subsampling on the non toxic comments since the number of non-toxic comments was very high than the number of toxic comments. Figure 2 shows our analysis of the frequency of toxic comments targetting each separate identity. As per the figure, most toxic comments are targetted towards white and black identities in our dataset.

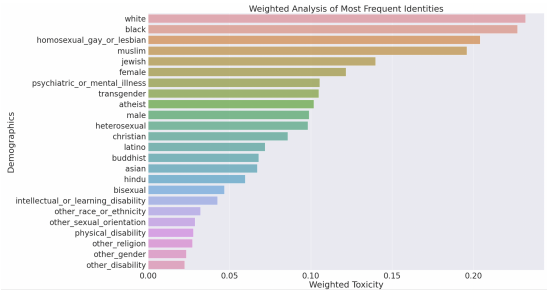

Figure 2: Weighted analysis of the most frequent identities present in the dataset.

Figure 3 shows a word cloud representing the 15 most frequent words occurring in comments that were tagged as of being about 'Asian' identity.

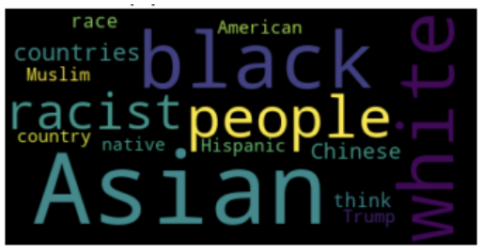

Figure 3: Word cloud representing the frequently occurring from the identity Asian

5.1 Identifying Protected Words

As per our initial hypothesis, we had to find the protected words which could lead to a biased classification. We followed the following process to identify the protected words. First, we grouped the identities from the dataset into 4 main groups as shown in Table 2

Group	Protected Identities
Religion	atheist, buddhist, christian, hindu, jewish, muslim
Race	asian, black, latino, white
Gender	female, male, transgender
Sexual Identity	bisexual, heterosexual, homosexual_gay_or_lesbian

Table 2: Grouping of Protected Identities

We then found the top 20 most occurring words in each group, and calculated the ratio of these words occurring in toxic comments and non toxic comments using the following formulae -

$$Toxic\ Ratio = \frac{No.\ of\ toxic\ comments\ with\ word}{total\ toxic\ comments}$$

$$Non - Toxic\ Ratio = \frac{No.\ of\ non{-}toxic\ comments\ with\ word}{total\ non{-}toxic\ comments}$$

We believe that the difference between these ratios will show us how prevalent the word is in toxic comments as against non toxic comments. If (Toxic Ratio - Non Toxic Ratio) > 0, then there is a chance that this word is a protected word since it occurs more frequently in toxic comments as compared to non toxic comments, which could make it a spurious feature for the toxicity model.

After calculating these ratios, we ranked the words based on these differences. We also manually identified whether these highly ranked words are relevant or irrelevant to the group, since we did see that many irrelevant words such as 'people' were present. We can see the top few such words for each group in Table 3.

We believe that the relevant words we identified this way put together would be our protected words. However, we did notice that some words related to less occurring identities like 'hindu' do not feature in this list since they do not have high ratios, however, we do believe that these should be protected words as well. Thus, we found the most frequently occurring words for each such less-occurring identities and added them to the list of protected words as deemed necessary, along with some filtering of irrelevant words. We found the following protected words as described in Table 4. along with their total occurences in our dataset.

We then used this data and our protected words and trained Standard Learning Without Z, Standard Learning With Z, and MIM models.

The main challenges we faced in training models on this data were that since the data is text data, after applying Count Vectorizer on the data, we were dealing with sparse data since there were a lot of features to deal with (about 150k). To

Identity Groups	Protected Attributes	Toxic Ratio	Non-Toxic Ratio	Ratio Difference	Relevance
	people	14.42	12.23	**2.19**	**Irrelevant**
	terrorist	1.53	0.56	**0.97**	**Irrelevant**
	trump	1.72	0.82	**0.89**	**Irrelevant**
	muslim	0.47	0.11	**0.35**	**Relevant**
religion	religion	0.99	0.70	**0.28**	**Relevant**
	islam	0.25	0.07	**0.18**	**Relevant**
	christian	0.20	0.07	**0.13**	**Relevant**
	jew	0.12	0.06	**0.06**	**Relevant**
	religious	0.79	0.73	**0.05**	**Relevant**
	atheist	0.10	0.06	**0.04**	**Relevant**
	white	4.97	1.37	**3.60**	**Relevant**
	racist	3.92	0.74	**3.18**	**Relevant**
	black	3.11	0.85	**2.26**	**Relevant**
race	people	14.42	12.23	**2.19**	**Irrelevant**
	trump	1.72	0.82	**0.89**	**Irrelevant**
	asian	0.09	0.05	**0.03**	**Relevant**
	latino	0.02	0.07	**0.01**	**Relevant**
	white	4.97	1.37	**3.60**	**Relevant**
	people	14.42	12.23	**2.19**	**Irrelevant**
	man	12.43	11.01	**1.41**	**Relevant**
	women	2.92	1.66	**1.25**	**Relevant**
gender	gay	1.47	0.30	**1.17**	**Relevant**
	trump	1.72	0.82	**0.89**	**Irrelevant**
	woman	1.71	0.89	**0.82**	**Relevant**
	right	8.39	7.69	**0.70**	**Irrelevant**
	male	1.37	0.76	**0.61**	**Relevant**
	transgender	0.28	0.10	**0.18**	**Relevant**
	sex	3.84	1.39	**2.44**	**Relevant**
	people	14.42	12.23	**2.19**	**Irrelevant**
	sexual	2.24	0.79	**1.45**	**Relevant**
	gay	1.47	0.30	**1.17**	**Relevant**
sexual_identity	right	8.39	7..69	**0.70**	**Irrelevant**
	homosexual	0.51	0.12	**0.38**	**Relevant**
	heterosexual	0.11	0.03	**0.07**	**Relevant**
	bisexual	0.02	0.01	**0.01**	**Relevant**
	percent	0.35	0.71	**-0.35**	**Irrelevant**
	men	23.89	25.58	**-1.70**	**Irrelevant**

Table 3: Ranked frequently occurring words for each group

overcome this, we then decided to let go of words/features that occur less than 10 times in the whole data.

6 Results

6.1 Standard Learning With Z

We trained a Logistic Regression Model and a Naive Bayes Model with the entire dataset. The Naive Bayes model did not perform as expected, so we decided to use the Logistic Regression Model as our Baseline Model. We got the following results for this model -

1. Accuracy of the model : 79.45%
2. Percentage of non toxic comments in test having protected words and have been wrongly classified as toxic: 9.72%

We wanted to see the influence of our protected words in the toxicity prediction. We categorized our protected words and then found the average of the coefficients for these groups. The groups are shown below in Table 5 -

We see the average coefficients for these groups in Table 6. We see that the protected words under the sexual identity category have high average coefficient value, which could mean that these words impart high bias to the toxicity label.

Protected Groups	Average Coefficients
religion	0.20
race	0.28
gender	-0.05
sexual_identity	0.43
misc	0.06

Table 6: Average Coefficients over groups for Baseline LR Model

Protected Words	No. of appearances in dataset
asian	305
atheist	199
bigotry	727
bisexual	61
black	7027
buddhist	61
catholic	3282
chinese	792
christian	2663
church	4950
gay	3557
heterosexual	294
hindu	60
homosexual	861
islam	1651
jew	237
jewish	997
latino	122
male	1836
man	5122
men	5249
military	1047
muslim	3449
priest	618
race	1703
racism	2177
racist	3843
religion	1911
religious	1789
sex	2772
sexual	2686
straight	652
transgender	484
transgendered	188
white	12822
woman	3339
women	9044

Table 4: No. of appearances of each identity word in the table.

6.2 Standard Learning without Z

For this fairness method, we removed the identified protected words from the comment text and trained a Logisitic Regression Model on this data. We got the following results -

1. Accuracy of the model : 78.1%
2. Percentage of non toxic comments in test having protected words and have been wrongly classified as toxic: 9.03%

We see that the results are not very different from the Baseline Model, and hence believe that simply removing the protected words from the dataset is not enough, there could be some proxies of these protected words which still influence the toxicity score.

6.3 Marginal Interventional Mixture model

We then used MIM to see if this would help remove some of the bias in the toxicity model. We used the FAX-AI library (FaX-AI, 2021) to train MIM model with this dataset. However, we saw that using the entire list of 37 protected words was causing issues due to computational power crunches, and decided to run it on a slightly smaller and more specific set of protected words first. These are the protected words we used -
Protected words = 'gay', 'black', 'islam', 'christian', 'white', 'muslim', 'homosexual', 'chinese', 'bisexual', 'jew', 'catholic', 'hindu', 'transgender', 'latino.
We got the following results -

1. Accuracy of the model : 78.79%
2. Percentage of non toxic comments in test having protected words and have been wrongly classified as toxic: 7.83%

We see that the percentage of misclassified comments has reduced, which shows that the model has become more fair.

7 Conclusion

We found that some words contribute to unintended bias while identifying toxic comments. We found a list of these protected words relating to different identities across the dataset. We then ranked these protected words according to our own metric which we found

Protected Groups	Protected Words
religion	religion, religious, atheist, buddhist, christian, hindu, jewish, muslim, catholic, church, islam, jew, priest
race	asian, black, latino, white, chinese, race, racism
gender	female, male, transgender, man, men, trangendered, woman, women
sexual_identity	bisexual, heterosexual, gay, homosexual, sex, sexual, straight
misc	bigotry, military

Table 5: Grouping of Protected words

would be essential in finding which words would contribute to bias the most.

We trained Standard Learning With Z, Standard Learning Without Z and Marginal Interventional Mixture models on this data using the identified protected attributes. We found that the accuracies of the models remain almost same. However, Standard Learning without Z doesn't help much since the percentage of non toxic comments classified as toxic also doesn't decrease much.

MIM seems to be more fair since it was able to reduce the percentage of non toxic comments wrongly classified as toxic. This indicates it was able to get rid of some biases in the model.

For future work, we believe that using a bigger list of protected words would reduce the bias even further. To test this, we would like to compare MDE values for the different models.

8 Contributions of group members

- Arpita Singh: Finding protected words, training models, generating graphs, analysis of results.
- Arundhati Gorkhe: Ranking protected words, training models, grouping model coefficients by identities, analysis of results.
- Arunima Sundar: Preprocessing, developing initial baseline models, finding model coefficients.
- Nandhinee Periyakaruppan: Preprocessing, identity analysis and EDA.

9 Source Code

Our project files and source code are uploaded here : Source code files

References

Alyssa Lees, Vinh Q. Tran, Y. T. J. S. J. G. D. M. L. V. (2022). A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv:2202.11176*.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (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.

FaX-AI (2021). Fair and explainable ai. *https://github.com/social-info-lab/FaX-AI*.

Google and Jigsaw, Perspective API (2017). . *https://www.perspectiveapi.com/*.

Ji Ho Park, Jamin Shin, P. F. (2018). Reducing gender bias in abusive language detection. *arXiv:1808.07231*.

Reichert, E., Qiu, H., and Bayrooti, J. (2020). Reading between the demographic lines: Resolving sources of bias in toxicity classifiers. *arXiv preprint arXiv:2006.16402*.