Nutritional Labels for Automated Decision Systems: The Effect of Word Embeddings on Bias

Ankush Jain
New York University
Brooklyn, NY
ankush.jain@nyu.edu

Vishnu Thakral
New York University
Brooklyn, NY
vvt223@nyu.edu

May 11, 2020

Abstract

Social interaction platforms are now a quintessential part of our lives as they have evolved into a resourceful space for sharing, learning, interacting, and marketing (SLIM) with huge crowd. Antagonistic to the good side of making the reach to a wide audience across various demographics, they sometimes become the ground for arguments and toxic comments. As a result, negative comments that are generally based on the demographics and social identities of a group end up being falsely associated with ”being toxic”. The ADS explores the influence of different word embeddings on unintended bias. In this work we benchmark the use popular word embeddings, including GloVe, FastText, ConceptNet NumberBatch, Custom trained Keras Embedding embedding and Meta Embeddings. The ADS uses a simple LSTM along with various combinations of word embeddings to give predictions.

1 Background

In recent times, we have seen the emergence of a new public interaction space in the form of social media platforms, discussion forums, and e-commerce platforms. These spaces have become a quintessential part of our lives by providing a constructive environment for better networking, or collective opinion and quorum based decision making experience. However, with the diverse ideas and demographics of the social media users, one does encounter the occasional disagreements and arguments. These can range from, rude or disrespectful commentary to outright toxic behaviour from a certain subgroup of users. At its worst, it has been observed that, the users tend to make negative comments based on the demographics and social identities[1] of their peers such as race, sex, orientation, nationality, ethnicity, religion, disabilities and more. As a result, the names or terms related to certain social identities tend to get falsely associated with “being toxic”.

This problem has been acknowledged by the scientific community and their efforts to address the crisis can be seen by the volumes of ongoing research on toxicity and social-sentiment analysis[2]. In this work, we are performing the nutritional labeling of a Natural Language Processing (NLP) and Machine Learning (ML) based Automated Decision System (ADS). In general, these NLP based systems make use of word embeddings for feature extraction and they are prone to such unintended bias of falsely associating the demographics rich terms with ”being toxic”. However, this ADS claims to minimize the unintended bias and we are going to benchmark its performance by producing nutritional labels for the ADS model as well as the word embeddings.

1.1 Goals of the Automated Decision System

The ADS benchmarked in this work is called **The Effect of Word Embeddings on Bias**[3]. It was developed as a solution to the Kaggle competition called **Jigsaw Unintended Bias in Toxicity Classification**[4], in July 2019 by the Conversation AI research group. The primary goal of the ADS is to develop a machine learning model to fairly identify toxicity in online conversations across a diverse range of conversations[5]. Toxicity was defined as ”anything rude or disrespectful or otherwise likely to make someone leave a discussion”.

In a previous iteration of the competition, the Conversation AI group had asked the Kaggle community to develop a Toxicity Classifier. After a rigorous analysis of the produced solutions, they realized that the toxicity models learned to falsely associate the terms related to the frequently attacked identities with toxicity[5]. For instance, the comment *”Continue to stand strong gay community. Yes,indeed, you'll overcome and you have.”* was misclassified as toxic. The statistical explanation to such unfortunate faux-pas is the fact that the users at the data source had overwhelmingly attacked certain identities. As a result, the models trained on this data had learned to broadly classify all the comments with the terms associated to these identities as toxic.

The numerous submissions for this competition allowed the development of the Perspective API. A research[6] on the nutritional labelling of machine learning models (called model cards) also generated a label for the Perspective API to showcase the concept. However, in our work, we are scrutinizing one of the ADS developed in this competition and not the actual API. The data set used in this ADS is composed of natural language text data. A common practice adopted for feature extraction from such data is to employ word embedding generation techniques or use pre-trained word embeddings. The secondary goal of this ADS is to present a comparative study of the effect of the various word embedding techniques on the fairness in identifying toxicity across a diverse range of conversations. The secondary goal of this ADS will be scrutinized in this work to identify the extent to which the different word embedding techniques achieve fairness.

2 Input and Output

The data for this ADS (and Kaggle competition) was collected from an open archive created by the Civil Comments platform right before it was shutdown in 2017. This was done to aid researchers in improving the civility in online conversations. The sponsors of this competition, Conversation AI, decided to extend this data set by adding annotations about the toxicity of the comments against different social identities. It was accomplished by the help of human raters.

The data set consists of the input attribute **comment_text** in text data type and the label **target** (toxicity) in float data type. But, the Conversation AI group has specified to perform thresholding at $target >= 0.5$ for interpreting toxicity in boolean data type. There are 24 auxiliary attributes (called Social Identity Attributes in this work) that provide the information about the attacked social group in the comments. But, the competition specified the analysis of the 9 identities which are primarily the most attacked identites.

The total size of the data is 1.9 million. The annotations about the toxic attributes/subidentities is available for a total of 42.6k comments from each social group. But, labels are available for only 1.8 million data and annotations for 40.5k data. Thus, the labelled data is split into 80:20 train-test ratio. Each social group is uniformly represented in terms of count. We will use the annotations for validating the fairness of the produced results. Annotations are not used as input features.

Table 1: Annotated/Non-Empty Data Distribution for Train and Test Data

TrainDataLength	AnnotatedTrainPerIdentity	TestDataLength	AnnotatedTestPerIdentity
1804874	405130	97320	21293

2.1 Data Profiling : Input Attribute (TEXT)

Table 2: Input Attribute Profiling

Input Attribute	EmptyCells	NonEmptyCells	Unique	Length			
				Min	Max	Mean	Median
comment_text	0	1804874	1780823	1	8813	297.23	202

The input attribute **comment_text** has zero empty values and it ranges from 1 to 8813 in length. The high range of input text and high variability of length must be properly handled by the feature extraction technique and the machine learning model.

2.2 Data Profiling : Output Attribute / Label (BOOLEAN)

Table 3: Output Attribute Profiling

Output Attribute / Label	EmptyCells	NonEmptyCells	UniqueVals	TrueVals	FalseVals
target	0	1804874	2	144334	1660540

The output attribute **target** (toxicity) has zero empty value, two unique values (0:False or 1:True) and a True:False ratio of 0.08:0.92. Hence, 92% of the training data is Non-Toxic (toxicity is false) and 8% data is Toxic (toxicity is true).

2.3 Data Profiling : Social Identity Attributes (REAL NUMBER)

Table 4: Social Identity Attributes Profiling

Social Identity Attribute	Unique	Max	Min	Mean	Median	StdDev
male	242	1	0	0.108687	0	0.267894
female	204	1	0	0.12767	0	0.305384
homosexual_gay_or_lesbian	124	1	0	0.025611	0	0.143739
christian	160	1	0	0.095268	0	0.256671
jewish	120	1	0	0.017863	0	0.122145
muslim	138	1	0	0.04946	0	0.202459
black	127	1	0	0.034393	0	0.1679
white	151	1	0	0.05695	0	0.21596
psychiatric_or_mental_illness	94	1	0	0.012083	0	0.089183

Social Identity Attributes or the annotations will be used in validating the fairness of the model predictions. All the above metrics have been calculated for 405130 annotated data entries (refer table 1). The value for each social group attribute in the data set denotes the probability of the comment being toxic and offensive to the particular social group. The probability will range from 0.0 (not toxic) to 1.0 (extremely toxic).

We can also observe that certain social identities such as homosexual_gay_or_lesbian, black, muslim, white, christian, male, female have high standard deviation from mean (in increasing order). We can expect high variation in toxicity against these social identities.

2.4 Data Distribution : Social Identity Attributes

To study the distribution of data across the various ranges of probability, we have divided the data into 5 buckets ranging from 0.0 to 1.0 at an interval of 0.2. We plot the frequency of data samples that fall in each probability range or bucket to observe the distribution of probabilities across the various social identities.

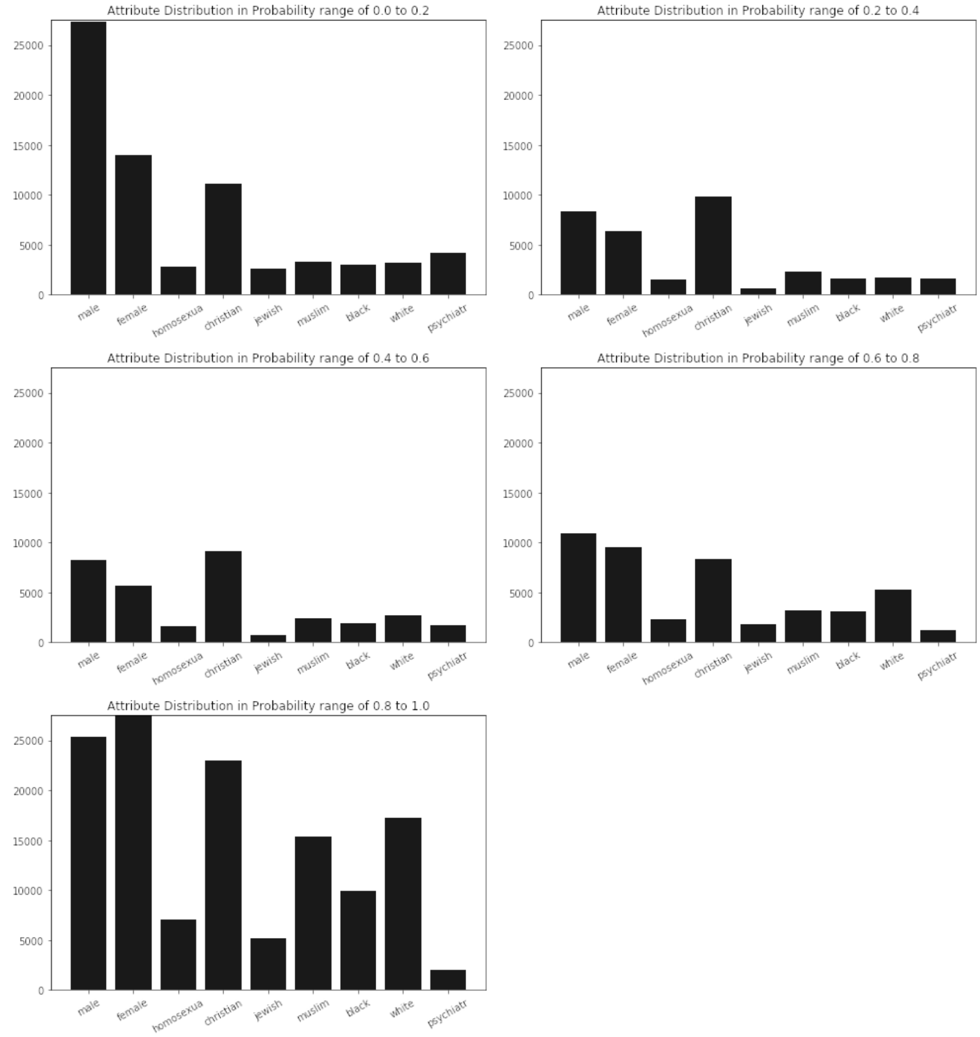

Figure 1: Data distribution for buckets of size 0.2

From the above distribution plots, it is evident that the social identities most represented in the toxicity dataet are female, male, christian and muslim. From the social identity attribute profiling, it was also evident that female and male identities had high means for the probability of toxicity. These plots have validated the results.

2.5 Data Correlations

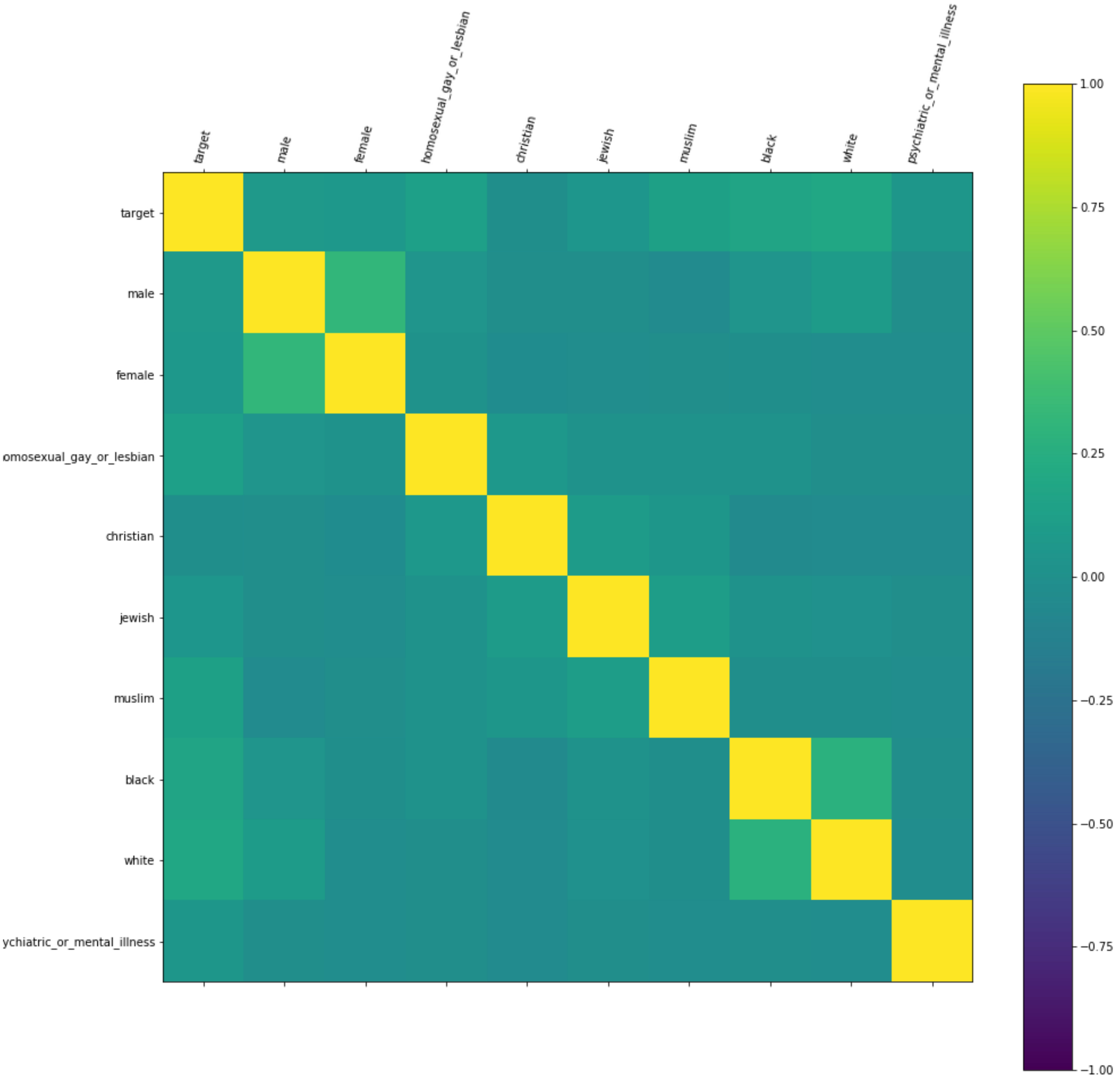

Figure 2: Correlation of social identity parameters with Target

After analysing the correlations plot, we can conclude with the following observations about the various social identities and target in dataset:
1) There is correlation between muslim and target
2) There is correlation between homosexual_gay_or_lesbian and target
3) The whites are correlated with blacks
4) The females are correlated with males
5) The christians are least correlated and the whites are most correlated with the target

3 Implementation and Validation

3.1 Data Cleaning and Pre-Processing

The ADS preprocesses the input text by converting the text into lower case followed by replacing short hand words like isn't to is not. After the mapping is done the text is cleaned by removing all special characters like ?!@' and additional white spaces. The ADS converts the target variable into categorical form by using threshold of 0.5 as the limit thus assigning toxicity = Yes(1) or No(0) instead of probabilities. The data is split into train and test while ensuring that the splits retain the identity labels to benchmark subclass. The final step in preprocessing is tokenization of words using .

3.2 High level Description of the Implementation

Toxicity classification is a sub-problem of text classification, which falls under the umbrella of sequence classification problems. The ADS uses the Keras implementation of the Long Short Term Memory (LSTM) architecture of Recurrent Neural Networks (RNN). Due to the sequential nature of the text data and the variation of the input text length across data samples, the application of LSTM is justified for this ADS. LSTM architecture is widely used in the various sequence classification problems as it addresses the long-term dependency problem with the general RNN for long text samples. An elaborate profiling of the text data has been presented in the next section which can further validate the usage of LSTM in the ADS.

The ADS also compares the effect of the different word embedding techniques on toxicity classification fairness across social identities and identify the fairest word embedding. It explores the data set trained word embeddings, the pre-trained embeddings (GloVe, BERT, Numberbatch and FastText) and meta embeddings by combining the different pre-trained embeddings.

3.3 Developer's Validation of the ADS

The ADS uses Area Under Curve (AUC) metric for benchmarking the bias across the various social identities and the word embeddings.

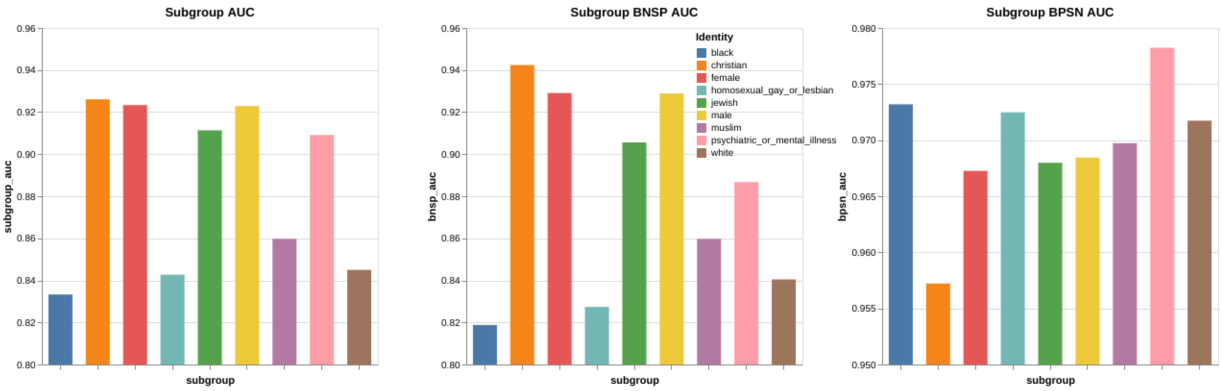

Figure 3: Sub-Group AUC plots for Bias Validation

Subgroup AUC: The AUC score for the complete subgroup/identity. A low score here implies

that the model fails to distinguish between toxic and non-toxic comments that mention this subgroup/identity.
BPSN AUC: Background Positive Subgroup Negative. A low value here means the model confuses the non-toxic examples that mention the subgroup with the toxic examples that do not.
BNSP AUC: Background Negative Subgroup Positive. A low value here means that the model confuses toxic examples that mention the subgroup with the non-toxic examples that do not.

After studying the AUC plots, we can see that te Subgroup AUC ranges from 0.80 to 0.93. This score is very good for AUC and it implies that we have good classifications. However, the ADS has room for improvement as it exhibits low AUC for certain classes like blacks and homosexuals.

4 Outcomes

4.1 Effectiveness Analysis of the ADS across Sub Populations

The overall accuracy for the ADS improves with GloVe and FastText word embeddings. The custom trained Keras Embeddings using the LSTM are the worst performing.

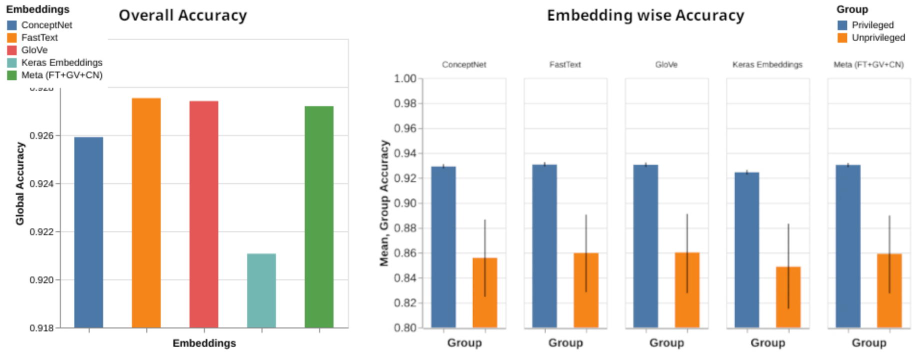

Figure 4: Embedding wise accuracy on Privileged/Unprivileged Groups

All the embeddings have pretty much good relatable accuracy with Keras Embedding having the least among all. In general all the embeddings have a lower accuracy for the unprivileged group thus pointing to the fact that we have more False Positive Rate for unprivileged class being labeled as toxic.

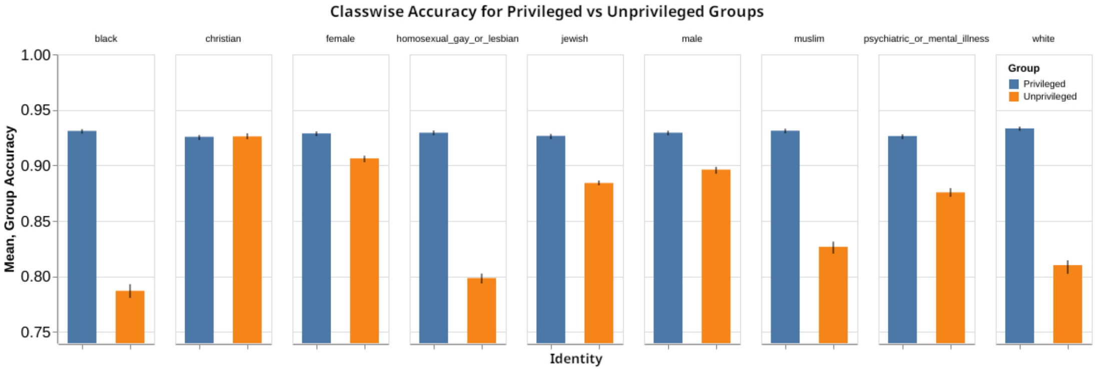

Figure 5: Social Identity wise accuracy on the Privileged/Unprivileged Groups

It is evident that the privileged group have a better accuracy in prediction of the toxicity label.

Thus unprivileged group are often classified wrongly and have a high false positive rate for being labeled as Toxic. Further, the black, homosexual, muslim and white sub populations have the lowest accuracy after classification.

4.2 Fairness/Diversity Measures to Benchmark the ADS

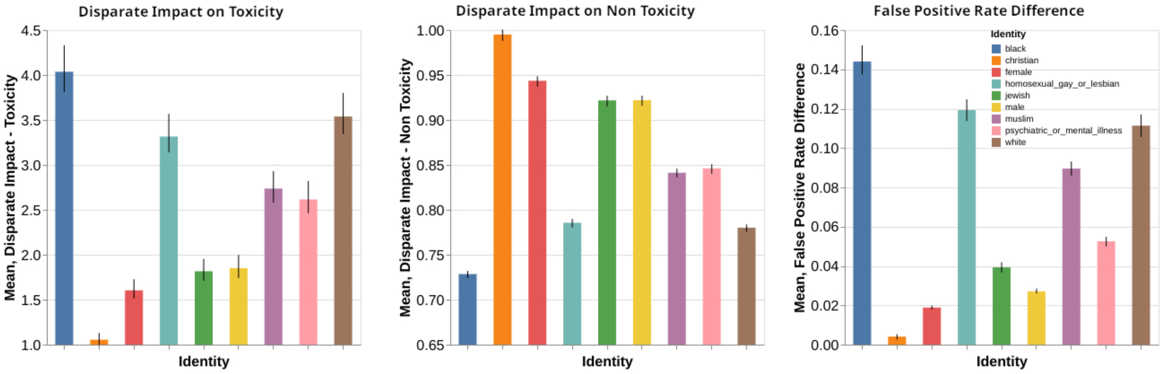

Figure 6: DI vs FPR on social identity attributes

Disparate Impact

Disparate Impact ≥ 0.8 is an indication that the unprivileged group has a higher probability of being labeled toxic. This is the case with unprivileged group in each category and justifies the assumption that negative comments that are generally based on the demographics and social identities of a group end up being falsely associated with "being toxic".

Figure 6 Plot 2 is a better indication of the same as we wish to have more Non toxic label predictions. Christian has a DI of 1. Interestingly Black and White races have low DI for Non Toxicity, which points to the fact that comments/text gets toxic when any race is involved, with a little higher magnitude for black race.

We, now consider a simple case of Gender Discrimination in employment scenario. With gender as protected class and employment as the target variable, we see that Female/Unprivileged group has less probability of getting a positive outcome (employement=Yes) in comparison to the privileged class i.e. male in the datasets displaying gender discrimination. In such a scenario, employment is a favourable outcome that we wish to increase for the unprivileged group. However, in our ADS, the Toxicity, which is denoted by 1 is not the outcome that we wish to increase. Infact, the purpose of the ADS is to minimize toxicity.

Ideally, a positive outcome in the ADS are marked non toxic comments, our assumption is that minority group will receive more toxic comments due to various reasons like being disliked by majority, being outnumbered and thus possible victim for mockery. If positive outcome is taken as toxic = Yes, the minority (unprivileged class) will receive more toxic comments and thus will result in a DI value more than 1. This is also supported by the results from the DI calculator.

A better approach would be to calculate DI for toxic=No as the positive outcome.

$$DI = \frac{Pr(Toxic = No|D = unprivileged)}{Pr(Toxic = No|D = privileged)}$$

False Positive Rate Difference

The False Positive Rate Difference acts as a really good fairness metric in case of this ADS. It gives us information about the samples within a sub group that are falsely predicted as toxic. The sole purpose of this ADS is to reduce the false positive cases, where a non toxic comment is misclassified as toxic.

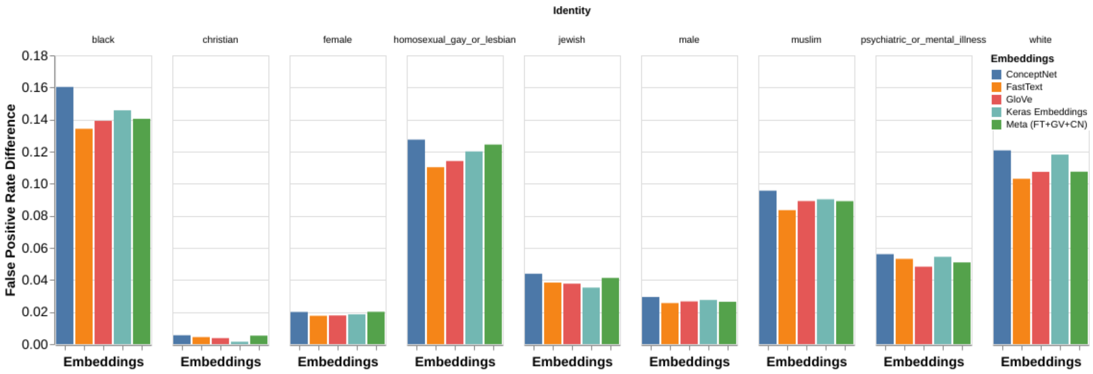

Figure 7: FPR comparison of embeddings on social identity attributes

ConceptNet Embedding has more False Positive rate for all the social identity attributes, closely followed by Keras Embeddings. FastText has the best overall accuracy as it produces the least FPR, which is important in our case. Having high FPR could be dangerous as it would add in to negative association of the unprivileged group with toxicity, something that we are trying to reduce.

Christian class has very low FPR difference thus meaning it has almost no involvement in labeling a text/comment as toxic or not. White and Black both race have high FPR difference thus hinting that in general the words related to these races have false associations with toxicity. Similarly, high FPR is also displayed by the homosexual clas..

4.3 Fairness/Diversity Measures to Benchmark the Word Embeddings

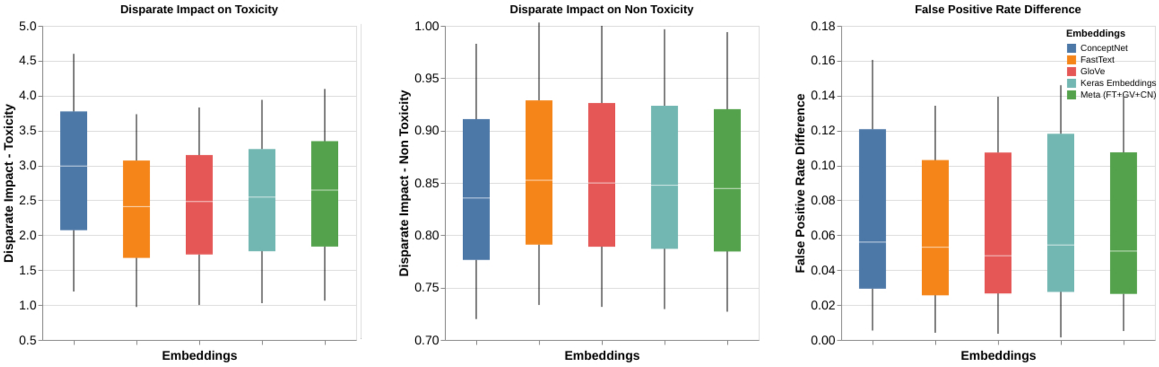

Figure 8: DI vs FPR on Embeddings for toxicity

All the embeddings have a high Disparate Impact with Toxicity(≥ 1), thus confirming that the text with tag related to unprivileged class is more probable to be labeled toxic. If target is taken as non-toxic the FastText ADS has the best DI and would be a good choice so as to reduce the misclassification.

False Positive Rate is a good measure in our case as we can have an estimate of how many instances are being falsely being associated to being toxic. A lower FPR is preferred and FastText provides the least FPR difference among classes. Keras Embeddings and Conceptnet embeddings have the worst FPR.

4.4 Explanation for Misclassification

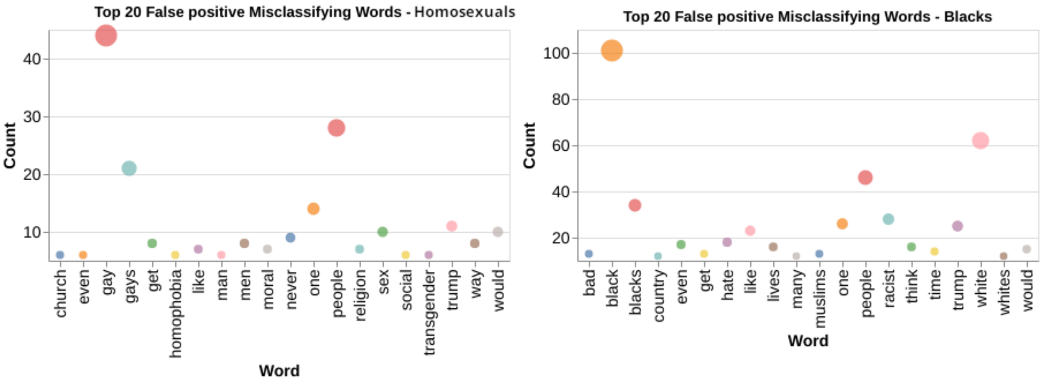

Figure 9: Top 20 FP Mis-classifying Words

In an effort to understand the cause for misclassification of the worst FPR Difference sub populations i.e. Blacks Homosexuals, we performed an analysis of the most common words. We found out all the words that are common among the black and homosexual false positive subsamples across the results generated by the 5 embedding techniques. The samples were consistently false positive misclassified by all 5 word embedding techniques.

Words like black, white, people, trump, racist etc. cause the FP misclassification of black population comments as toxic. And, similarly, words like gay, people, one, trump, sex etc. cause FP misclassification of homosexual population comments as toxic. Just the bear presence of these words can cause the misclassification.

5 Summary

During our analysis of this ADS, we have determined that the dataset and the provided features within the dataset have been beneficial at determining the quality of the ADS and the embeddings. We were able to determine the various fairness metrics with the help of the auxiliary features regarding the social identities of the commentors.

When analysed with other fairness metrics like False Positive Rate Difference and Disparate Impact, we saw strong inverse correlation between the [Subgroup AUC, Subgroup BNSP AUC] and our fairness metrics. Although, BNSP AUC and BPSN AUC consider the background data into and not just the subgroup data, they performed very well in determining fairness. Hence, the author's decision to benchmark their work on the basis of subgroup AUC was a good idea. The ADS has managed to generate high accuracy while maintaining general fairness across the predictions.

Although, most fairness metrics and sub-population accuracies have very minor differences, we cannot ignore the fact that some sub populations (namely blacks, homosexuals, muslims and whites) are more biased towards getting a non-toxic comment missclassified as toxic. Hence, these stakeholders are not going to receive the levels of fairness as compared to other sub populations.

The fairness metrics of our choice i.e. Sub Population Accuracies, Disparate Impact and False Positive Rate Difference are beneficial to the fairness regulatory bodies, the members of affected sub groups and the internet users in general will benefit from our work. We have introduced transparency and some explainability about the ADS and the various word embedding techniques. The ADS gave the fairest results with FastText and GloVe embeddings.

An average accuracy of correctly classifying toxic comments is very respectable. So, this ADS can be considered for deployment. However, its unfairness towards blacks, homosexual people, muslims and whites must be addressed. We did observe disparate impacct for non toxic comments to be in th range of 75% to 95% and this is a solid starting point for building a fair Toxicity classification ADS.

To improve upon this solid foundation, certain improvements can be made to the system. The training data must be increased in volume and it should be uniformly divided. During the preprocessing analysis, we observed that classes like male, female and christian had higher volumes of training data. The very fact that these classes had higher fairness score signifies the importance of having uniform and high volumes of training data for each sub population.

Lastly, the word embedding like FastText and GloVe showed better fairness as compared to ConceptNet and the meta embeddings. If these embeddings are trained on better datasets, then they will give fairer results. Keras Embeddings were obviously compromised because they were trained on this dataset. And our dataset had unbalanced representation for certain classes.

References

[1] Turner, John; Oakes, Penny (1986). ”The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence”. British Journal of Social Psychology. 25 (3): 237–252. doi:10.1111/j.2044-8309.1986.tb00732

[2] Mika V.Mäntyläa, Daniel Graziotin Miikka Kuutilaa. ”The evolution of sentiment analysis—A review of research topics, venues, and top cited papers”, Computer Science Revie Vol 27. https://doi.org/10.1016/j.cosrev.2017.10.002

[3] Kaggle, “Jigsaw Unintended Bias in Toxicity Classification”, https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/

[4] Kaggle, “The Effect of Word Embeddings on Bias”, https://www.kaggle.com/nholloway/the-effect-of-word-embeddings-on-bias

[5] Jigsaw, Medium, ”Unintended bias and names of frequently targeted groups”, https://medium.com/the-false-positive/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23

[6]Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, ”Model Cards for Model Reporting”, https://arxiv.org/abs/1810.03993