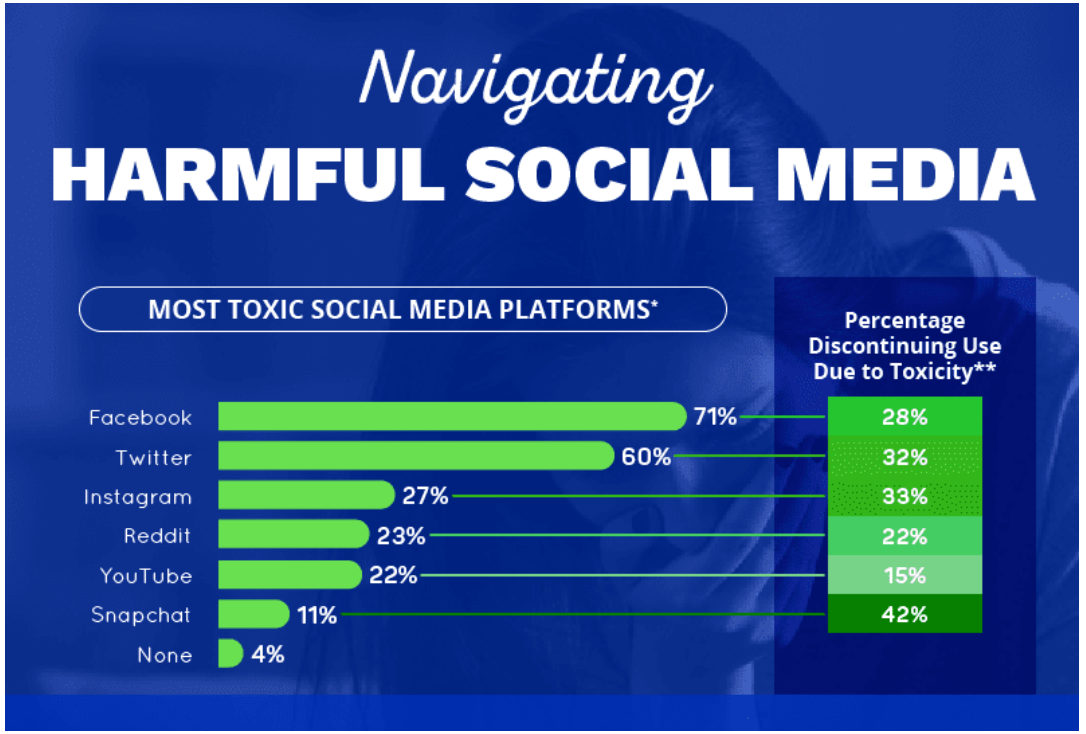
Nutritional Labels for Automated Decision Systems:

The Effect of Word Embeddings on Bias

By:
Ankush Jain
Vishnu Thakral

Background

- Are public interaction spaces like social media all good and platform for sharing, learning, interacting, and marketing?
- Could they unintentionally be promoting toxicity?



Source: https://www.digitalinformationworld.com/2019/11/building-credibility-on-social-media.html

ADS System

The Effect of Word Embeddings on Bias[1].
Developed as a solution to the Kaggle competition called **Jigsaw Unintended Bias in Toxicity Classification[2]** in July 2019 by the Conversation AI research group.

Goals:

1. Develop a machine learning model to fairly identify toxicity in online conversations across a diverse range of conversations. (anything rude/disrespectful or otherwise likely to make someone leave a discussion")
2. 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.

[1] "The Effect of Word Embeddings on Bias", https://www.kaggle.com/nholloway/theeffect-of-word-embeddings-on-bias
[2] Kaggle, "Jigsaw Unintended Bias in Toxicity Classification", https://www.kaggle.com/c/jigsawunintended-bias-in-toxicity-classification/

Data

- Source

open archive created by the Civil Comments platform right before it was shutdown in 2017.

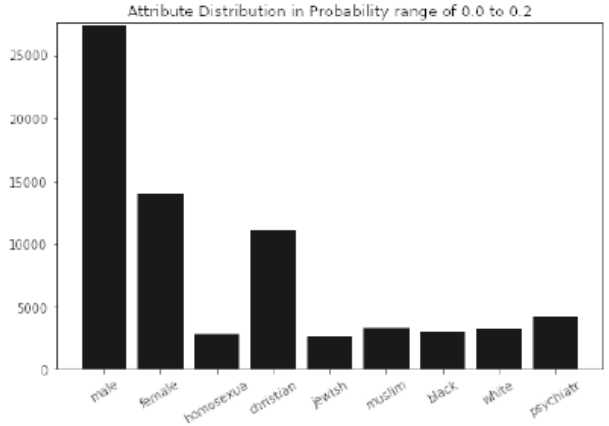
- Profiling

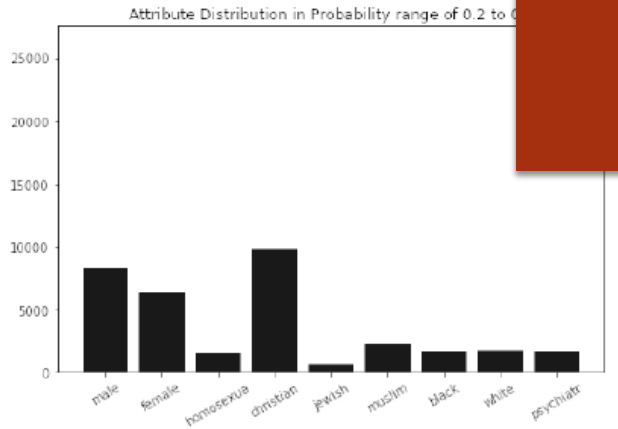
TrainDataLength	AnnotatedTrainPerIdentity	TestDataLength	AnnotatedTestPerIdentity
1804874	405130	97320	21293

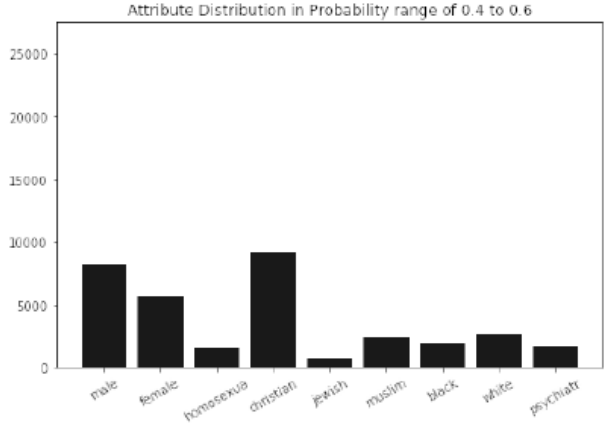
Output Attribute / Label	EmptyCells	NonEmptyCells	UniqueVals	TrueVals	FalseVals
target	0	1804874	2	144334	1660540

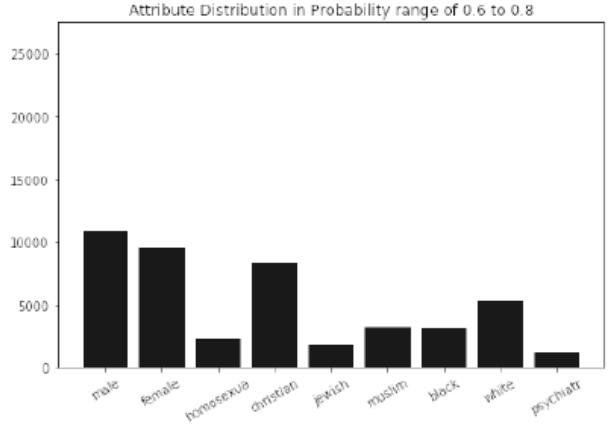
Social Identity Attribute	Unique	Max	Min	Mean	Median	StdDev
male	242	1	0	0.108687	0	0.267894
female	204	1	0	0.12767	0	0.305384
homosexual_gay_or_lesbian	124	1	0	0.025611	0	0.143739
christian	160	1	0	0.095268	0	0.256671
jewish	120	1	0	0.017863	0	0.122145
muslim	138	1	0	0.04946	0	0.202459
black	127	1	0	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

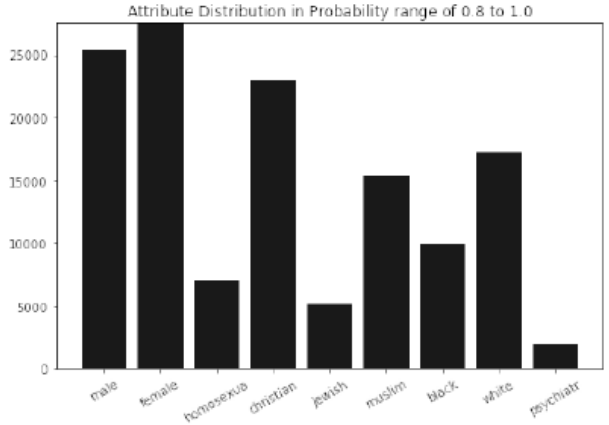
Data Visualization











Implementation of ADS

Toxicity classification is a sub-problem of text classification, which is under the umbrella of sequence classification problems.

The ADS preprocesses the input text by converting the text into lower case, replacing short hand words like isn't to is not removing all special characters like ?!@' and additional white spaces. Target is converted to categorical variable and words tokenized.

Embeddings used:

- ConceptNet
- FastText
- GloVe
- Keras Embedding (LSTM architecture of RNN)
- Meta (FastText + GloVe + ConceptNet)

Validation

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.

Validation

Area Under Curve (AUC)

This metric is used for benchmarking the bias across the various word embeddings.

Subgroup AUC

The AUC score for the entire subgroup, a low score means the model fails to distinguish between toxic and non-toxic comments.

Background positive, subgroup negative(BPSN) AUC

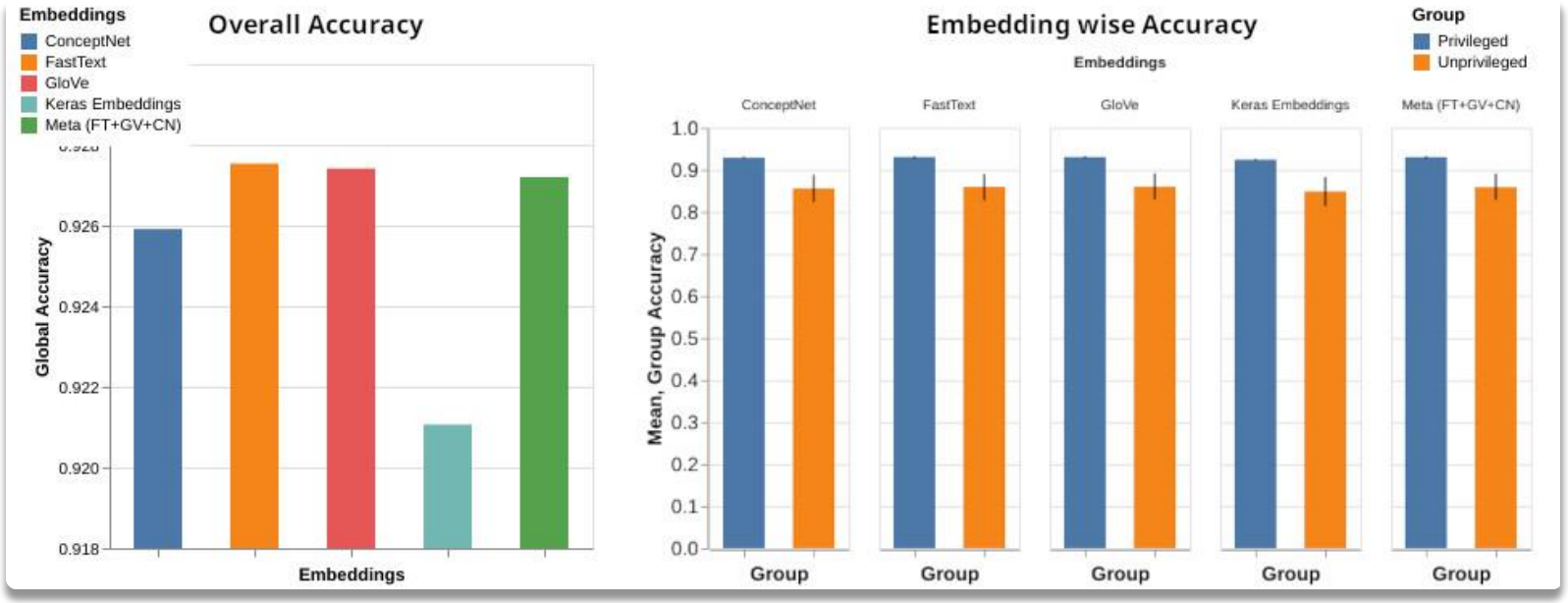
A low value of BPSN means the model confuses non-toxic examples of group with toxic that do not.

Background negative, subgroup positive(BNSP) AUC

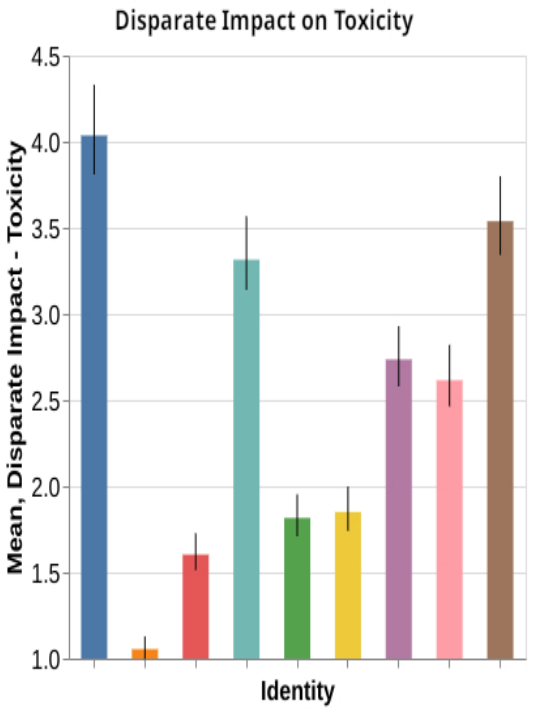
A low here means the model confuses toxic examples that mention the subgroup with non-toxic examples that don't.

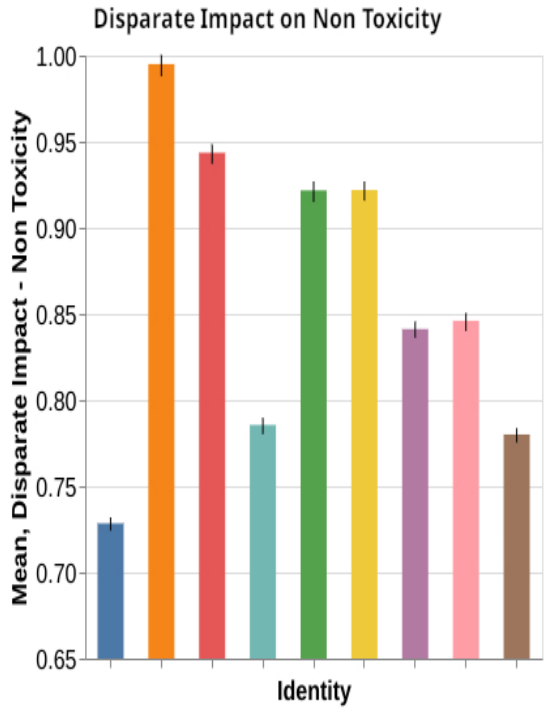
Outcome

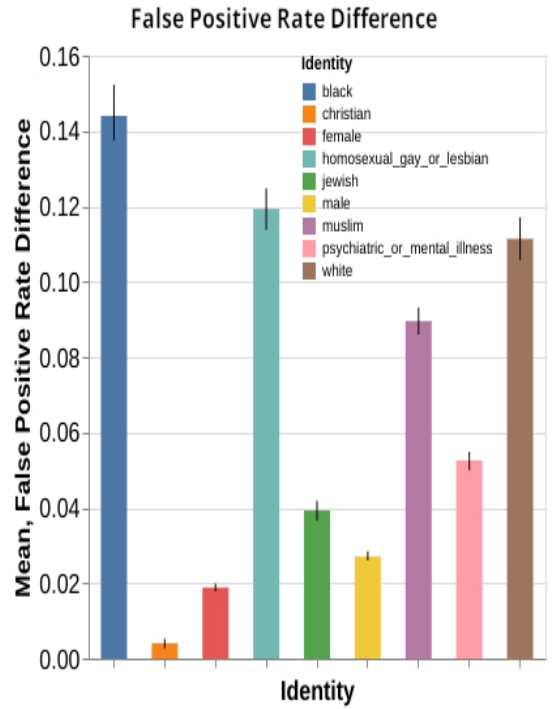
- All embeddings have pretty much good relatable accuracy
- All embeddings have a lower accuracy for the unprivileged group => more False Positive Rate for unprivileged class being labeled as toxic.



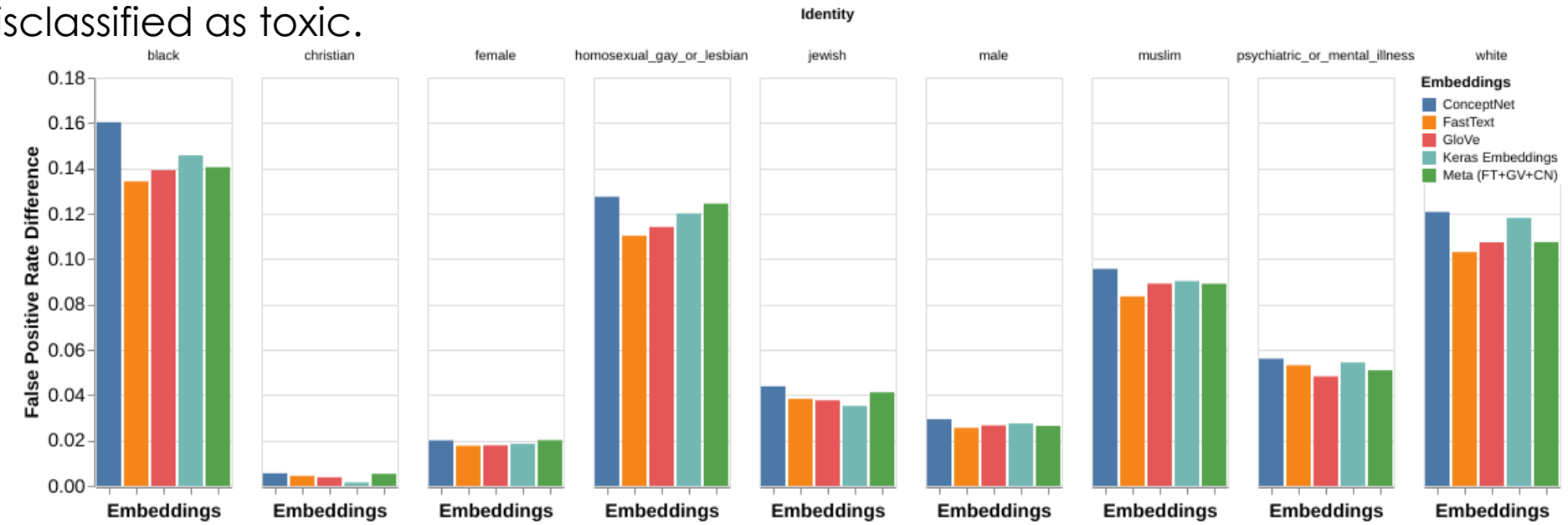
- Disparate Impact >1 is an indication that the unprivileged group has a higher probability of being labeled toxic.
- Christianity has a DI of 1.
- Black and White race both have low DI for non toxicity, i.e comments/texts get toxic when any race is involved, with a little higher magnitude for black race.



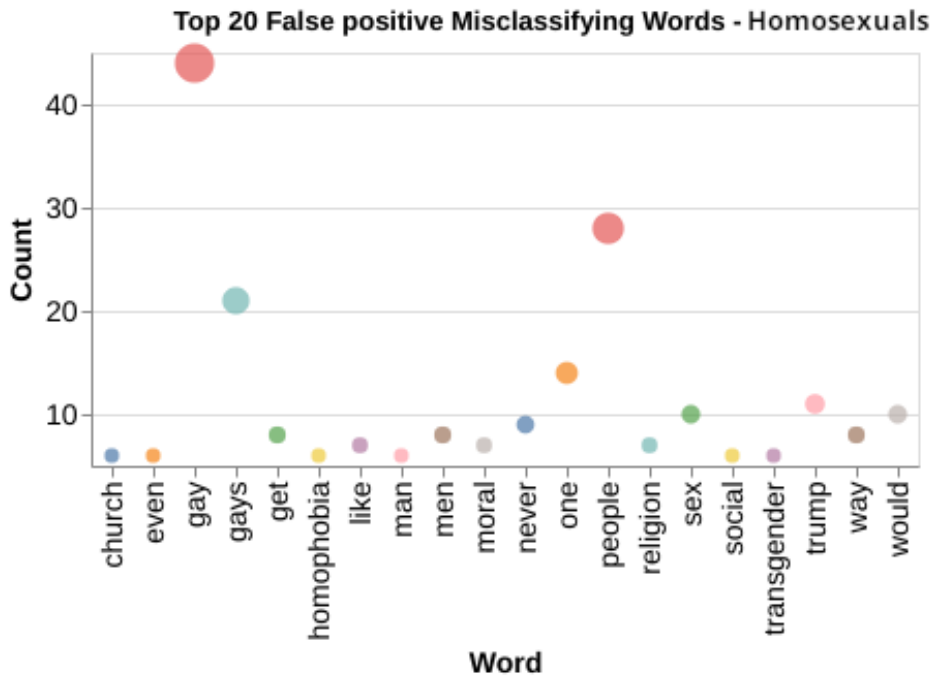


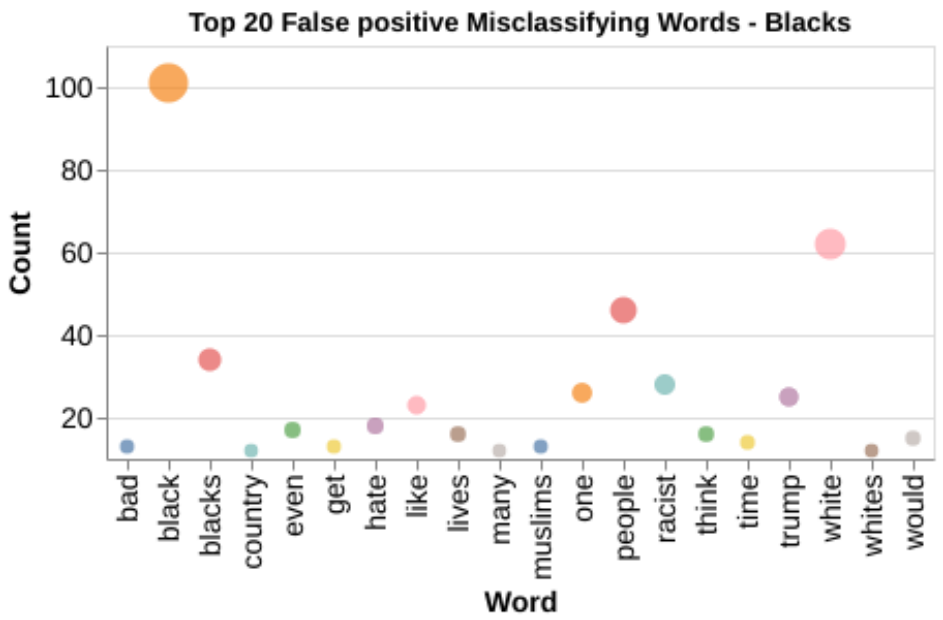


- The False Positive Rate Difference acts as a really good fairness metric in case of this ADS as it gives us information about the samples within a sub group that are falsely predicted as toxic.
- The sole purpose of this ADS is to reduce the false positive cases, where a non toxic comment is misclassified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



Misclassifications Explained





Summary

- The dataset had unbalanced representation for certain classes.
- 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 compromised because they were trained on this dataset.