**JIGSAW UNINTENDED BIAS IN TOXICITY CLASSIFICATION**

**Ahmad Alsharef Tanya Sharma**

ahmadalsharef994@gmail.com tanshu2708@gmail.com

KIIT University, Bhubaneswar BIT, Mesra

**Nidhi Chaudhary Ayush Pandey**

nidhichaudhary7075@gmail.com ayush.p321@gmail.com

Galgotias University, Noida LNCT, Bhopal

[[1]](#footnote-1) *Abstract*— In this paper, we have shown our approach of measuring Unintended Bias in Toxicity Classification. We illustrate how this approach can be used to evaluate various toxic comments and messages in various public forums and talk pages. We also demonstrate how imbalances in training data can lead to unintended bias in resulting models, and therefore potentially unfair applications. We have used a general supervised approach using

State-of-art algorithms and have used various external embeddings to improve our accuracy

Of the model. We demonstrate that this approach reduces the unintended bias without compromising overall model quality.

# INTRODUCTION

With the use of machine learning for a wide variety of tasks, researchers have identified unfairness in ML models as one of the growing concerns in the field. Many ML models are built from human-generated data, and human biases can easily result in a skewed distribution in the training data. ML practitioners must be proactive in recognizing and counteracting these biases, otherwise, our models and products risk perpetuating unfairness by performing better for some users than for others.

Our interest is in improving text classification models used to identify toxicity in comments from online discussions, but the evaluation methods presented here can be applied to a broad range of classification applications. “Toxicity”, defined as anything that is rude, disrespectful, or unreasonable that would make someone want to leave a conversation, is an inherently complex and subjective classification task. Machine learning systems, if not constrained, will often learn the simplest associations that can predict the labels, so any incorrect associations present in the training data can produce unintended associations in the final model. Toxicity models specifically have been shown to capture and reproduce biases common in society, for example misassociating the names of frequently attacked identity groups (such as “Gay”, and “Muslim” etc.) with toxicity. This unintended model bias could be due to the demographic composition of the online user pool, the latent or overt biases of those doing the labeling, or the very selection and sampling process used to choose which items to label.

In the following sections, we describe related work, then discuss a working definition of unintended bias in a classification task, and distinguish that from “unfairness” in an application. We then demonstrate that a significant cause of unintended bias in our baseline model is due to the disproportionate representation of data with certain identity terms andprovide a way to measure the extent of the disparity. We then propose a simple technique to counteract that bias by adding data. Finally, we present metrics for evaluating unintended bias in a model and demonstrate that our technique reduces unintended bias while maintaining overall model quality.

# Related Work

Various recent works have been published on defining how the concepts of fairness and unintended bias apply to machine learning models. Researchers have proposed various different matrices for the evaluation of fairness in models. (Kleinberg, Mullainathan, and Raghavan 2016) and (Friedler, Scheidegger, and Venkatasubramanian 2016) both compare several different fairness metrics. These works rely on the availability of demographic data about the object of classification in order to identify and mitigate bias. (Beutel et al. 2017) presents a new mitigation technique using adversarial training that requires only a small amount of labeled demographic data.

Very little prior work has been done on fairness for text classification tasks. (Blodgett and O’Connor 2017), (Hovy and Spruit 2016) and (Tatman 2017) discuss the impact of using unfair natural language processing models for real-world tasks but do not provide mitigation strategies. (Bolukbasi et al. 2016) demonstrates gender bias in word embeddings and provides a technique to “de-bias” them, allowing these more fair embeddings to be used for any text-based task. Our work adds to this growing body of machine learning fairness research with a novel approach to defining, measuring, and mitigating unintended bias for a text-based classification task.

### .

# Methodology

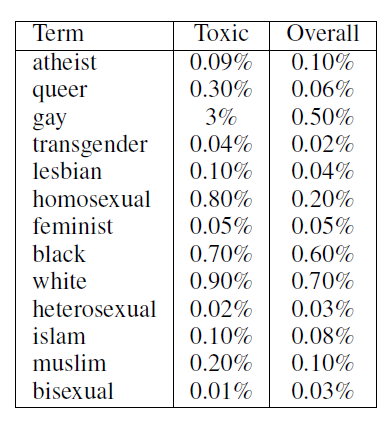
In this paper, we work with a text classifier built to identify toxicity in comments from various Talk Pages, public forums and social media sites. The model is built from a dataset of 1.8 million comments, each labeled by human raters as toxic or non-toxic. A toxic comment is defined as a “rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.” All versions of the model are convolutional neural networks trained using the Keras in TensorFlow.

**Pre-processing - Quantifying bias in the dataset**

Identity terms affected by the false-positive bias are disproportionately used in toxic comments in our training data. For example, the word ‘gay’ appears in 3% of toxic comments but only 0.5% of comments overall. The combination of dataset size, model training methods, and the disproportionate number of toxic examples for comments containing these words in the training data led to overfitting in the original toxicity model: it made generalizations such as associating the word ‘gay’ with toxicity. We manually created a set of 51 common identity terms and looked for similar disproportionate representations.

Table 1 illustrates the difference between the likelihood of seeing a given identity in a toxic statement vs. its overall likelihood.

**Term Toxic Overall**



**Table 1: Frequency of identity terms in toxic comments and overall.**

In addition to a disproportionate amount of toxicity in comments containing certain identity terms, there is also a relationship between comment length and toxicity, as shown in Figure1.



**Figure 1: Percent of comments labeled as toxic at each length containing the given terms.**

The models we are training are known to have the ability to capture contextual dependencies. However, with insufficient data, the model has no error signal that would require these distinctions, so these models are likely to overgeneralize, causing the false positive bias for identity terms.

**Model Details**

Initially, we have created a Simple LSTM(Long Short-Term Memory) model where the information flows through a mechanism known as cell states. In this way, LSTMs can selectively remember or forget things. The information at a particular cell state has three different dependencies. This model had an accuracy rate of 92%. After using the word-embeddings and genism library our model’s accuracy improved up to 93% on the same dataset.

**BERT Model**

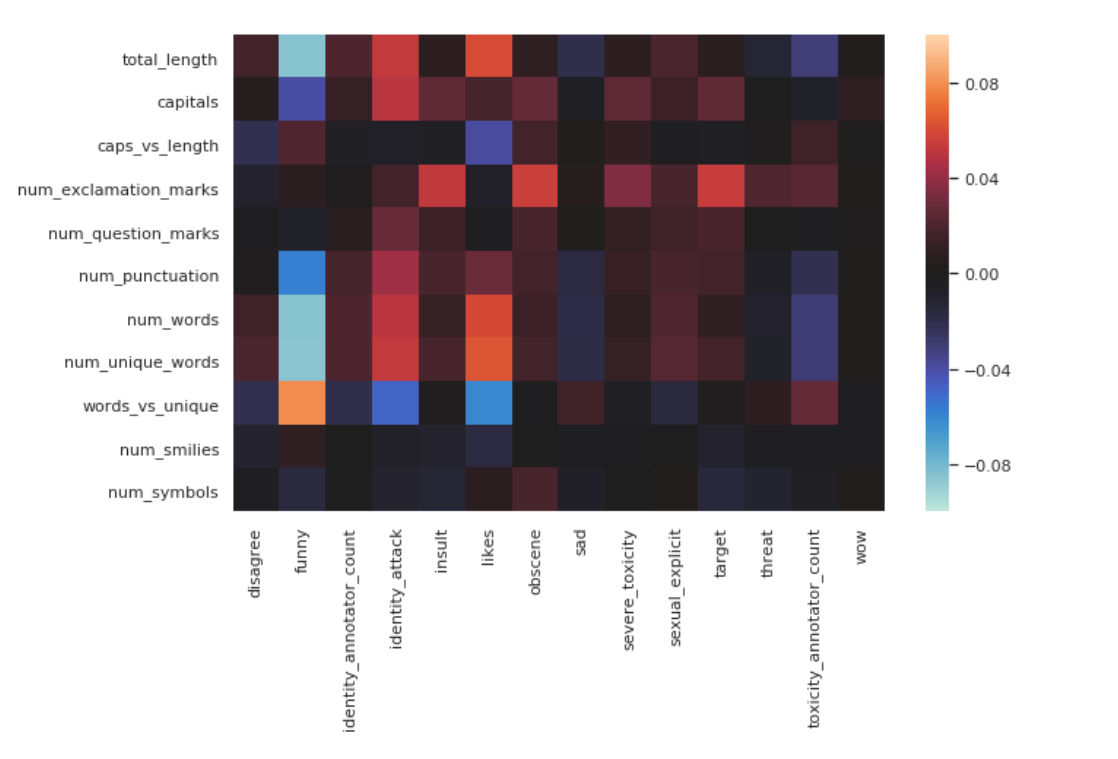
Finally, we implemented our model using BERT (Bidirectional Encoder Representations from Transformers) which proposed in a paper published by researchers at Google AI. BERT’s key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling. This is in contrast to previous efforts that looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The paper’s results show that a language model that is bi-directionally trained can have a deeper sense of language context and flow than single-direction language models. In the paper, the researchers detail a novel technique named Masked LM (MLM) which allows bidirectional training in models in which it was previously impossible. The BERT model gave an accuracy of 94%.

# Experimental Results

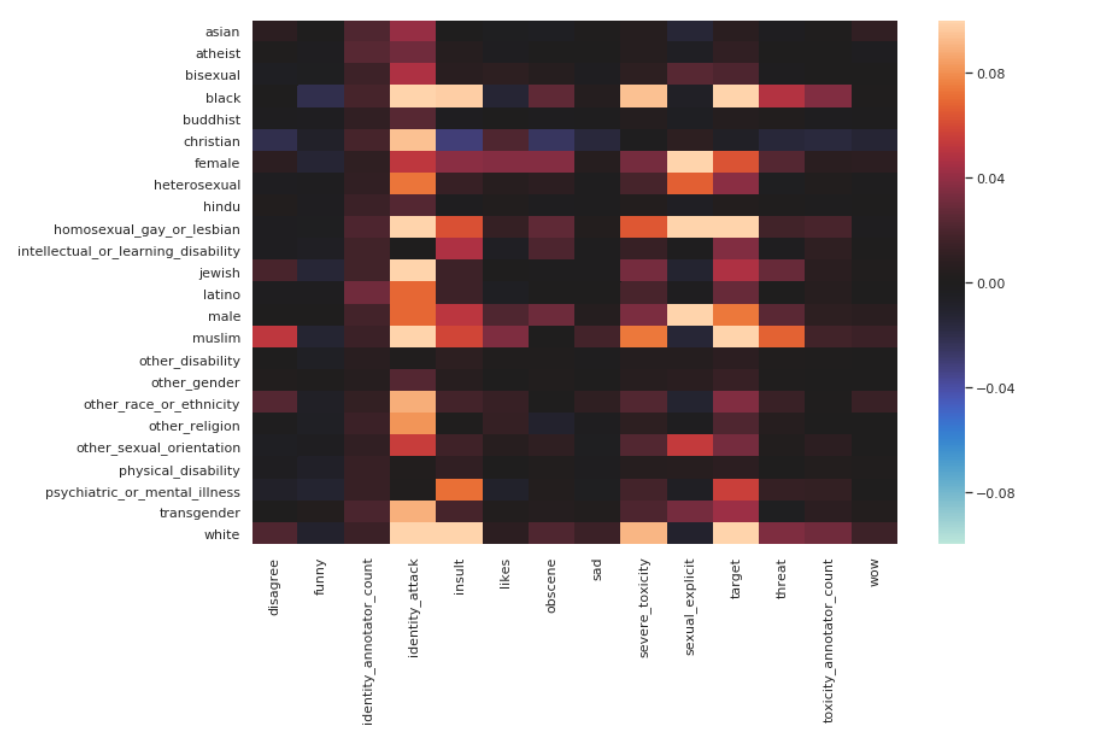
General Test Set

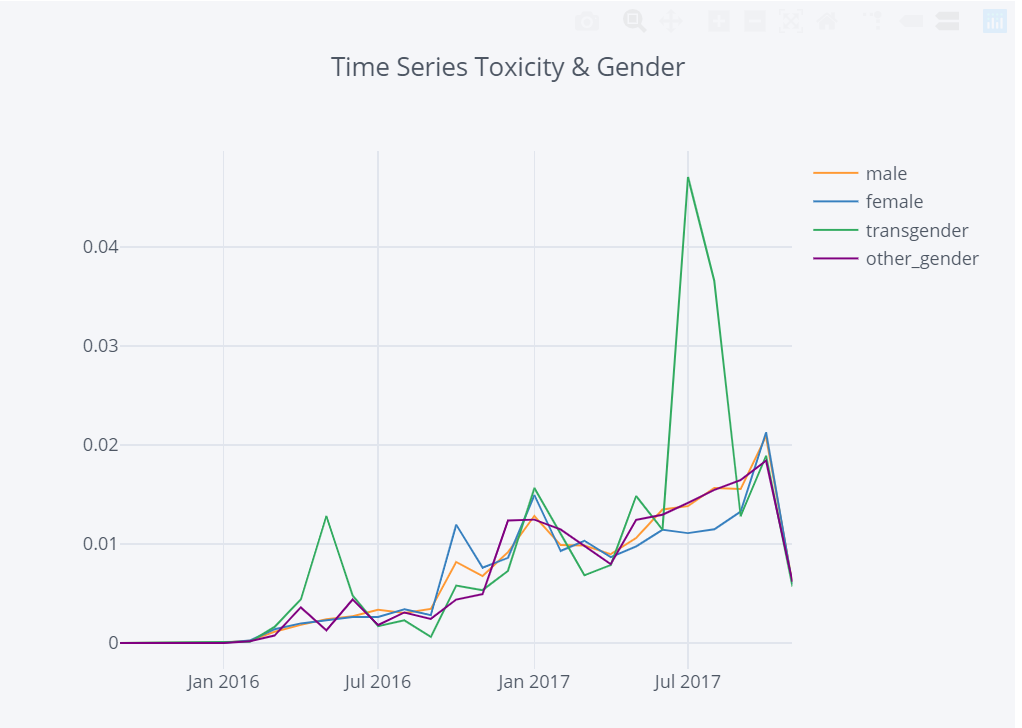
We use a general held out test set sampled from the original data set of Talk Page comments and various other online forums. This set evaluates overall model performance (intended bias), but it does not provide much information about unintended bias. We use this set to ensure that bias mitigation techniques do not drastically hurt overall performance.

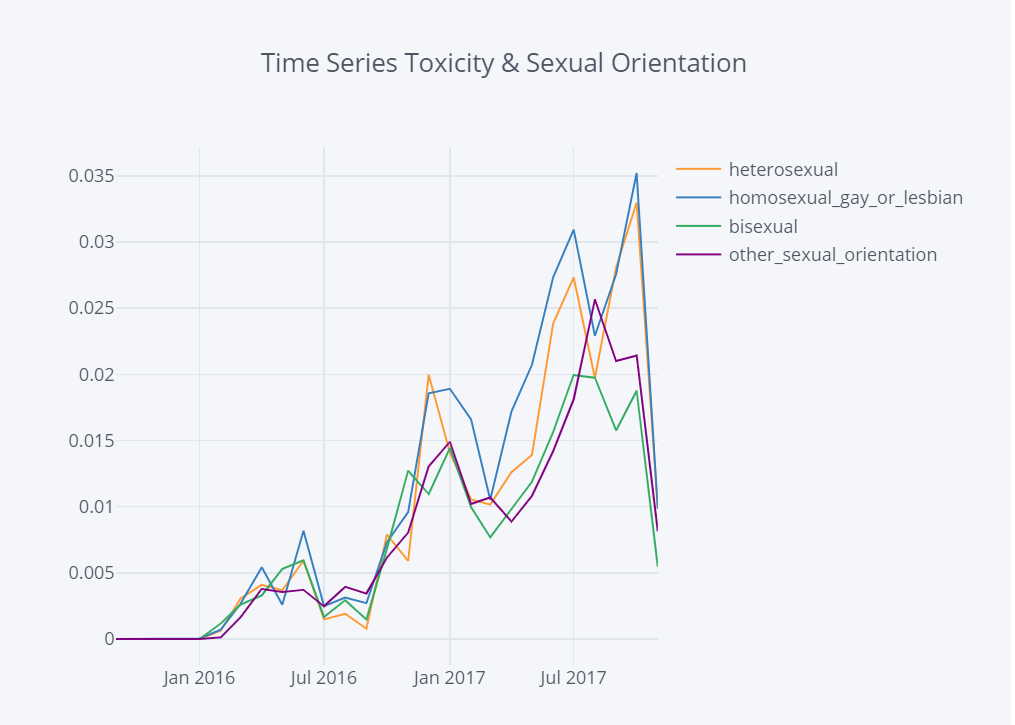
Correlations between new features and targets in heatmap:

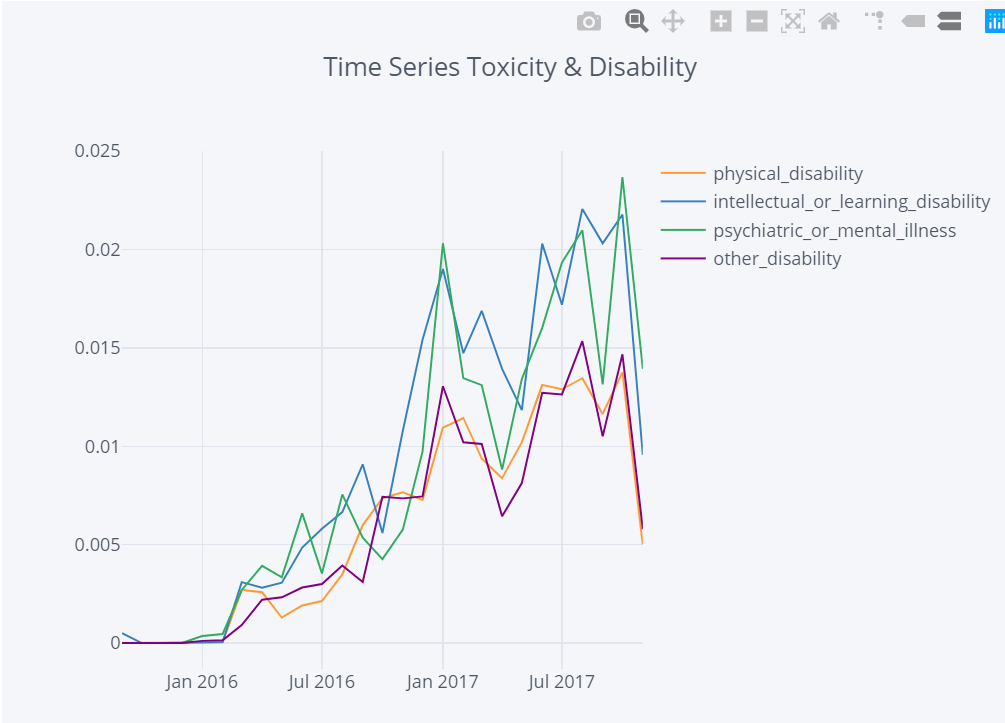


Percent of toxic comments related to different identities, using target and population amount of each identity as weights:

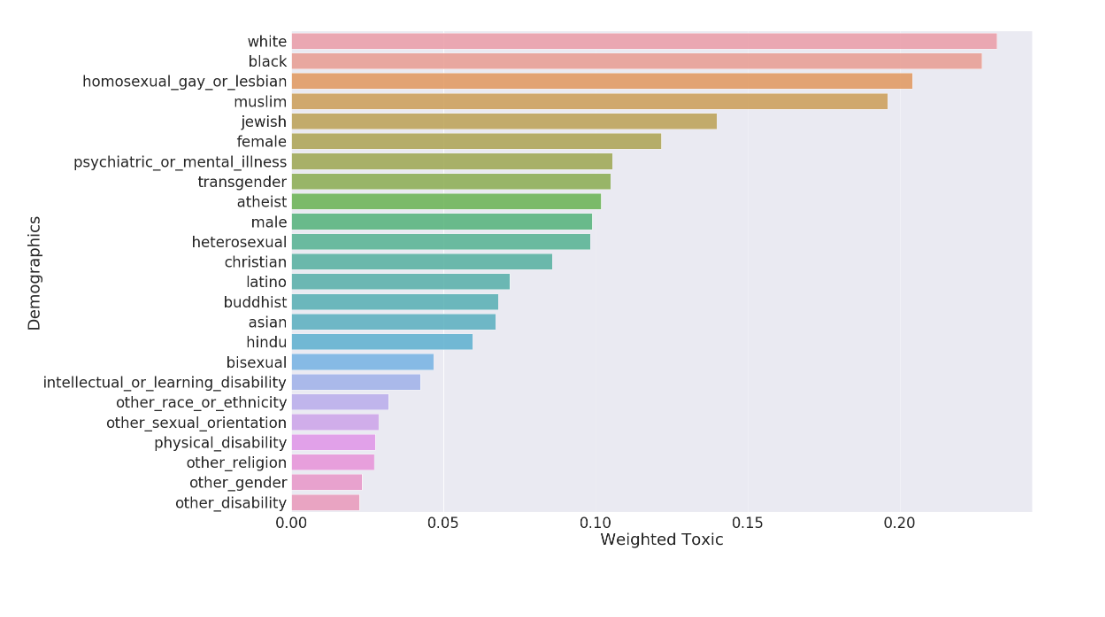


****

****

****

****

The correlations between identities and the comment labels

# Conclusion

In this paper, we have proposed a definition of unintended bias for text classification and have implemented various methods to improve the accuracy of the model. In this work, we presented multiple approaches for toxic comment classification. We find that a large source of errors is the lack of consistent quality of labels. Finally, we achieved a final accuracy of 94.179 using our proposed model.

References

1. Bert-Model:<https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>
2. Simple EDA Text Preprocessing: <https://www.kaggle.com/nz0722/simple-eda-text-preprocessing-jigsaw>
3. LSTM Model: <https://en.wikipedia.org/wiki/Long_short-term_memory>
4. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification: <https://arxiv.org/pdf/1903.04561.pdf>
5. Challenges for Toxic Comment Classification: An In-Depth Error Analysis: <https://www.aclweb.org/anthology/W18-5105>

1. [↑](#footnote-ref-1)