Dealing with Unintended Bias in Toxicity Classification

Prateek Malhotra
Department of Computer Science
University of California - Los Angeles
Los Angeles, CA 90025
prateekmalhotra@cs.ucla.edu

Tanmay Sardesai
Department of Computer Science
University of California - Los Angeles
Los Angeles, CA 90025
tanmays@cs.ucla.edu

Abstract

Machine learning approaches designed to deal with toxic comment classification on online forums are biased in the sense that they incorrectly learn to associate names of frequently attacked identities with toxicity. Our goal in this project is to come up with an approach to detect toxic comments while mitigating unintended model bias occurring due to an imbalance in the identities of people who were offended. We're attempting to solve this problem as part of an ongoing kaggle competition organized by Jigsaw which involves a specialized metric for measuring unintended bias and a dataset of comments with labeled identities of individuals.

1 Introduction

We will be participating in Jigsaw Unintended Bias in Toxicity Classification competition on Kaggle located at https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification. The competition has publicly available dataset which we will be using to build our model. The competition also has a private dataset which will be released on June 26th. As this is after our quarter ends we will not be able to report these final results.

2 Evaluation

For evaluation we will be using the metric used in the kaggle competition. The evaluation metric combines multiple ROC-AUC scores on different subsets of the dataset as described in Borkan et. al [2].

3 Expected Deliverables

For this project the expected deliverables will be the kernel that we turn in for the kaggle competition. Along with this we will also submit a video describing our work and some challenges that we faced in place of the final presentation.

4 Related Papers

In 2018, Jigsaw (an Alphabet Subsidiary) launched a Toxic Comment Classification challenge [5] and many of the submitted approaches were used to build an online API for identifying toxic comments in online forums [4]. Many developers and users working with this API noticed a bias in the way the alogrithm was assigning toxicity values as described in the competition FAQ [5] and so a new competition has now launched which relates to correct identification of toxic comments and removing unintended bias in the learned model.

In this project, we will attempt to solve this problem by building upon many recent approaches [3; 6] to deal with bias in text classification and other NLP tasks. Our starting point will be to build upon the Benchmark kernel [1] provided by the creators of this competition which achieves a score of 88.3 (with the current rank 1 having a score of 94.152).

5 Work division

We will be splitting the work equally. Both of us will be involved in reading related research papers and implementation of the project.

References

[1] Daniel Borkan. Benchmark Kernel - Unintended Bias in Toxicity Classification. 2018. URL: `https://www.kaggle.com/dborkan/benchmark-kernel`.

[2] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *arXiv preprint arXiv:1903.04561*, 2019.

[3] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. 2018.

[4] Perspective Google. Perspective API. 2018. URL: `https://www.perspectiveapi.com/`.

[5] Conversation AI Jigsaw. Toxic Comment Classification. 2018. URL: `https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview/description`.

[6] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.