DSGA-1017
Spring 2022

Project Proposal

Project partners: Chloe Zheng, Preston Harry

ADS: Detecting toxicity across a diverse range of conversations

Data: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

Data are taken from a Kaggle competition titled “Jigsaw Unintended Bias in Toxicity Classification”

The dataset contains comments and associated commenter data from an app, “Civil Comments”, meant to allow users to discuss news articles in a non-toxic way. The dataset includes a target indicating whether a comment is toxic, the text of the comment itself, and a range of demographic features about the commenter including gender, sexuality, religion, and race. There are also labels for the target to describe the type of toxicity displayed in a given comment.

Code: https://www.kaggle.com/code/dborkan/benchmark-kernel/notebook

Code is taken from a submission to the previously mentioned Kaggle competition titled “Benchmark Kernel”.

The script includes data preprocessing, text tokenization, training a CNN using Keras, and subgroup bias analysis using AUC. Given the large dataset, a random sample of the dataset will be used to accelerate training.

This ADS attempts to predict the toxicity of comments for which the labels may depend on pre-existing biases or subjective experiences of those labeling the data. An algorithm may exacerbate these biases by disproportionately assigning toxic labels to certain groups, causing disparate treatment in comment regulation. Additionally, if the public conversational data could be used to trace a person’s identity, this could negatively impact their employment opportunities or endanger their lives if they identify as a protected class.