Responsible Data Science (DS-UA 202) Final Project

Project Partners: Mimi Chen, Kieren Gill

ADS: Toxic Comments Classifier

Project Proposal

The competition we have decided to analyze is the Toxic Comment Classification Challenge, its aim is to identify and classify toxic online comments. The datasets were obtained from kaggle.com's Toxic Comment Classification Challenge[1], and the specific code implementing the ADS was obtained from user @jagangupta 's submission for this contest.

We selected this specific ADS because there has been an alarming increase in the amount of hate speech on the internet. Online hate speech in the UK and US has risen by 20% since the start of the pandemic, according to a new report[2]. We believe that by identifying toxic comments and offensive phrases, users and developers alike will be be able to engage in better discussion online and promote a healthier environment. This ADS will use a training dataset of Wikipedia comments that have been labeled by human raters as different types of toxicity: toxic, severe_toxic, obscene, threat, insult, and identity_hate. Using this dataset, the contestant created a model that predicts the probability of each type of toxicity for each comment.

In relation to the topics discussed in this course, the ADS we have chosen is susceptible to various biases. It is susceptible to pre-existing biases because of the biases that human raters have when labeling the comments, emergent biases, and technical biases as well. This ADS also relates to the transparency and interpretability module of this course. We hope to explore how the

[1] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/
[2] Baggs, Michael. "Online Hate Speech Rose 20% during Pandemic: 'We've Normalised It'." BBC News, BBC, 15 Nov. 2021, https://www.bbc.com/news/newsbeat-59292509.

ADS model makes its decisions, and how we can explain the ADS in interpretable terms to a wide audience.