Springboard–DSC

Capstone #2 Project Proposal: Toxic Comment Classification Challenge

Report by Zohreh Asaee

April, 2022

The main goal of this project is to help online discussions become more productive and respectful.

Problem Statement:

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) is working on tools to help improve the online conversation. One area of focus is the study of negative online behaviours, like toxic comments (*i.e.* comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (*e.g.* some platforms may be fine with profanity, but not with other types of toxic content).

In this project, a multi-headed model will be built that is capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. A dataset of comments from Wikipedia's talk page edits will be used.

Context:

As mentioned earlier, a multi-labelled model will be built to detect the toxicity. The dataset is available on the Kaggle website. It consists of a large number of Wikipedia comments labelled by human raters for toxic behaviour. The types of toxicity are toxic, severe toxic, obscene, threat, insult, and identity hate. The model will predict the probability of each type of toxicity for each comment.

Criteria for Success:

The dataset contains training and test sets. The criteria for success is labelling the training dataset as accurately as possible.

Scope of Solution Space:

The texts will be classified into the six groups of toxicity (as used for labelling the test dataset).

Data Sources:

Kaggle
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data