Springboard–DSC

Capstone #2 Project Proposal: Toxic Comment Classification Challenge

Report by Zohreh Asaee

April, 2022

Summary:

In this project, we will build classification models with the goal of predicting the level of toxicity of comments that have been previously labelled to reflect such levels.

The underlying main goal is to help online discussions become more productive and respectful.

Problem Statement and Context:

The threat of abuse and harassment on social media platforms is a serious concern. These platforms struggle to effectively facilitate conversations and may lead to limiting or completely blocking 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 for this team is the study of negative online behaviors, like toxic comments (*i.e.* comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). However, 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, models will be built that are capable of computing the likelihood associated with 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, the link to which is included below.

Scope and Criteria for Success:

The business problem will be modelled as a multi-class classification problem, where the classes are given by the different levels of toxicity. The inputs to the models will be comments and the outputs will be probabilities associated with each of the classes.

Multiple models will be compared and evaluated with respect to appropriate performance metrics, and interpretability analyses will be attempted. We anticipate that some of the classes might be imbalanced, and therefore we might need to use appropriate techniques to handle this.

The project will be limited to the categories defined in the dataset, and we might need to collapse some of them to make the problem less imbalanced.

Data Source:

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