Capstone project 2: Toxic Comment Classification Challenge

The goal of the project:

To build a multi-headed model that is capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate better than Perspectives' current models.

Data:

A large number of Wikipedia comments which have been labeled by human rater for toxic behaviors are provided. The types of toxicity are:

Toxic, Severe_toxic, Obscene, Threat, Insult, Identity_hate

Research question:

Predicting a probability of each type of toxicity for each comments.

Link of data:

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

Potential audience：

Policy makers, regulators, parents, school officers