**PROJECT PROPOSAL**

Nowadays, more and more social media applications are leaning towards providing a safer and more all-inclusive environment for its users. A lot of backlashes was faced by the social media applications for explicit images, videos, and toxic comments towards the user. Hence, according to an article from social media today, Instagram is rolling out new moderation tools which will use machine learning to detect offensive language and spam. This could also mean that comments that are considered discouraging, ill-intentional, or outright hateful towards any user is being detected and removed.

Using Machine Learning, I have tried to implement a Toxic Comment Classifier using a data set that contains examples of both toxic, and non toxic comments. Using tokens that can preempt a comment being toxic or not, I am trying to increase the accuracy of an already existing project taken from “<https://www.kaggle.com/watermasterz/toxic-comments-classification/notebook>”.

This project makes use of four different models, and compares the four different models’ accuracy on how the comments have been classified. It uses neural networks and I will experiment with different values of parameters and by changing the configurations of neural networks using the knowledge from the course, that will make the model more accurate.

Based on these experiments that I will try, I will construct a model that is a more accurate model using the same methods that this project has used.

The main challenge that I will face in doing this is that the accuracy of the already existing project is about 95%. Any machine learning model that is close to being a perfect model has an implicit challenge already, that is, it is very difficult to increase the accuracy of these models even by a slight percentage. Using 1D convolution layers, 1D average pooling layers and densely connected regular neural network layers, this project tries to classify the toxic comments. We can use this machine learning model and build on it to create a toxic common classifier for any of the social media applications that are present in the market currently.