Group Proposal of Final Project
Cheng Zeng

- What problem did you select, and why did you select it?

  I used text classification to handle toxic and divisive content, which is a challenging problem for many major websites. Quora is a platform that enables people to learn from each other, where false premises, insincere question could be a big challenge. I selected Quora Insincere Questions Classification model to identify and flag insincere questions.

- What database will you use? Is it large enough to train a machine learning or different algorithms?

  I used the training and test dataset of "Quora Insincere Questions Classification" I found on the Kaggle. The dataset has over 200,000 records, which is large enough to train a machine learning.

- What type of neural network will you use? Will it be a standard form of the network, or will you have to customize it? What other machine algorithms you will be used?

  I used BERT (Bidirectional Encoder Representations from Transformers), which is a neural network-based technique for natural language processing pre-training. BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling.

- What software will you use to implement the neural network or different algorithms? Why?

  I used python to implement neural network. Python is a preferred programming language when it comes to text classification because of its simple syntax and number of open source available.

- What reference materials will you use to obtain sufficient background on applying the chosen algorithm to the specific problem that you selected?

  BERT is a breakthrough for NLP. It could train the language model based on entire set of words in query rather than the traditional way of training on the ordered sequence. It's brand-new, powerful, practical and has foreseeable future, so I chose it as the algorithm for text classification problem.

- How will you judge the performance of the network? What metrics will you use? Provide a rough schedule for completing the project.

I used the accuracy and F1 to judge the performance of the network. It considers both precision and recall of the test to compute the score.
Time schedule:
6.5-6.15 code building and debugging
6.16-6.18 report write-up
6.18-6.25 ppt making and wrapping up