Progress Report for Text Classification Competition: Twitter Sarcasm Detection

Oran Chan (wlchan2) and Edward Ma (kcma2)

Progress Made

- Completed data exploration and exploitation. Having 5000 equal distributed label training data. Split data to training set and evaluation set and sizes are 4500 and 500 respectively.
- Possible to find external data to enrich the dataset but considering the efforting of searching and data processing. We proposed to generate synthetic data. Generated different sizes of synthetic data for evaluation.
- Trained model based on the pre-trained neural network model (BERT and RoBERTa) and achieved a good result which exceeds the baseline.

Remaining Tasks

- Refractor coding for easier understanding
- Summarize the effectiveness of synthetic data
- Prepare the documentation and presentation about what we did and how does it work

Challenges

- Spent lots of time on model fine-tuning