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 instead of looking for external data as it involves lower effort. Generated different sizes of synthetic data for evaluation. From 0.5 times to 10 times.
- Evaluated deep neural network model architecture for building classification model
- 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. It includes the comparison among different sizes of synthetic data and models.
- Prepare the documentation. The focal point is how we can leverage synthetic data to boost up model performance with minimum human effort.
- Prepare the presentation material about what we did and how it works

Challenges

- Spends time on understanding the relationship between response and main thread content.
- Learn subword algorithms such as WordPiece (adopted by BERT) and Byte Pair Encoding (adopted by RoBERTa)
- Learn transformer architecture (i.e. the base architecture of BERT and RoBERTa models)
- As using transformer models, computation resource requirement is high. It takes several days to complete several epochs.