Student Information

- Name: 賴琮運
- Student ID: 112034571
- GitHub ID: ggbdmn
- Kaggle name: 112034571
- Kaggle private scoreboard snapshot: /img/pic0.png

Firstly, I tried to implement TFIDF on text and Random Forest as a baseline. Then, I looked deeper into the data and tried to preprocess it in a better way. Finally, I cleaned the text in several ways(mentioned in the code notebook), based on the data structure and characteristics. Also, I generated 4 types of features(W2V_on_text, TFIDF_on_hashtag, BERT_on_text, SentimentAnalyzer_on_text). I tried to utilize several models including RF, XGB, LSTM, and major voting through all, but LSTM lasted the best. Through the process, due to the large scale of the dataset, it is hard to precisely allocate the efficacy of RAM, CPU, GPU and even the cache storage, as I have no experience in doing this. However, after all did I find the way to activate my GPU and all other resources.

P.S. I do not think it is a good idea to append new description and new work to do especially 1 DAY before handing in the homework.