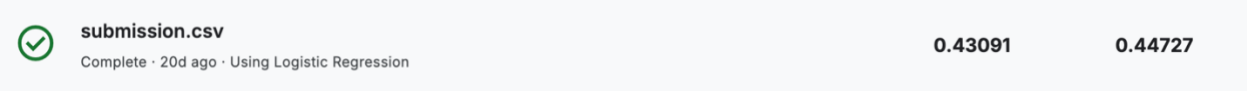
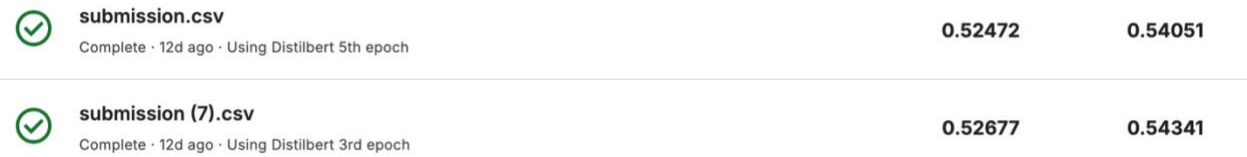
DM2024-Lab2-Homework

Name: 陳培熹
Student ID: 113065425
GitHub ID: Tedious8
Kaggle name: Tadeus

During the beginning of the lab 2 homework, I used traditional NLP techniques like combining TF-IDF and N-grams and logistic regression, and it got 44% accuracy. During this time I use several preprocessing techniques such as lowercasing, translating emojis, transforming slang words, lemmatization, etc.



At the mid phase, I try using a contextual embeddings approach using the DistilBERT model. The model shows good results by increasing its epoch until the third epoch, and then it becomes overfit. The accuracy I got is 54%, so it's quite a significant progress. Compared to the first approach (traditional NLP techniques), I didn't do any specific text preprocessing on the LLM model; I directly fed the data to the model because I believe that the text would be more meaningful and isn't necessary if doing the text preprocessing.



Finally, I try using a different model called CardiffNLP, and I need to train it until the fifth epoch, and then I stop before overfitting. The accuracy I got is 58%. Although it's not much progress compared to the mid-phase, in my opinion the result is still good.