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.

\odot	submission.csv	0.43091	0.44727
	Complete · 20d ago · Using Logistic Regression	0.43091	0.44727

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.

\odot	submission.csv Complete · 12d ago · Using Distilbert 5th epoch	0.52472	0.54051
\odot	submission (7).csv Complete · 12d ago · Using Distilbert 3rd epoch	0.52677	0.54341

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.

\odot	submission (5).csv	0.56090	0.58260
	Complete · 10d ago · Using Cardiffnlp 5th epoch	0.56980	