Pt.3 Model Report

Preprocessing

step1 - data loader: 根據所需欄位載入後，根據train data或是test data進行拆分，然後合併emotion欄位，進行初步的資料整理。
step2 - preprocessing: 對資料進行清洗，如進行小寫轉換，並利用nltk.corpus進行停用詞去除，讓資料更加簡潔，並將資料進行斷詞處理。將斷詞後的資料利用PorterStemmer和WordNetLemmatizer進行字根處理，留下最精要的部分。

Model test

model1 - w2v + random forest 先利用word2vec進行embedding，轉換好相關向量後提供分類模型進行使用(這裡記得載入模型以便日後使用和修改)。這裡的分類模型我選擇使用random forest classifier，然後節點分類用entropy進行評估。原本有嘗試使用小樣本加上SVM，但嘗試後效果不佳，然後試著用全部的樣本丟進SVM跑，但樣本太大，SVM運行太久完全不可行，故放棄。
model1 - hyperparameter 這裡我選擇用random search找出合適的超參數進行優化，這裡有另外對random search進行交叉驗證數和評估方法的設定。原本有嘗試用grid search進行超參數選擇，但同樣因為樣本數太大、運行時間太久故放棄。
model2 - transformer 我採用了Hugging Face 的 distilbert-base-uncased 模型進行文本分類。這個模型是一種精簡版的 BERT，而該Hugging Face 的AutoModelForSequenceClassification，可以設置情緒標籤數量對應的output layer節點。由於樣本量較大，為了加快訓練速度，我只訓練了 1 個 Epoch，但設置了每 50 步保存一次模型權重並進行評估，選擇準確率最高的模型作為最佳模型。但由於同樣運行時間過長，有待之後再進行改善和方法上的修正，所以總結我採取了model1。