Kaggle Competition

113065542 盧子涵
Kaggle 名稱: luzi8451

資料前處理:

1. 將 emoji 轉換成特定 token
   - ✧ 全部 emoji 皆轉為同一 token: '😂': '[emoji]'
   - ✧ 不同 emoji 轉為不同的 token: '😂': '[joy]'

   測試後的結果為轉為不同的 token 有較好的效果，後使用第二種方法
2. 移除空白等字元
3. 利用 data_identification.csv 中的資料拆分出 training set 及 testing set
4. 統一 submission 跟 testing data 的資料排序，以方便後續預測結束後填入
5. 利用 oneHot encoding 將標籤轉成編碼

訓練模型:

目的: 想利用不同的模型來達到計算資源及準確度的平衡，使用了 TinyBert, distilbert, BERT 做為測試模型

下圖為第一次提交的結果:

第二次提交使用了 TinyBert 作為模型、epoch=4:

第三次使用了 BERT 模型、epoch=3:

第四次使用了 distilbert 模型、epoch=5:

綜合以上所做的實驗，儘管 BERT 訓練的 epoch 較少，但其還是表現最好的，再來是 distilbert 和 TinyBert，但在訓練過程中 BERT 也是花費時間最多的(一個 epoch 需跑 1hr 左右)，若是資源有限、但想獲得不錯的結果也可退而求其次選擇較小的模型；另外在比賽結束後和同學討論發現也可使用在預處理時便使用 Tweeter 文本做訓練的模型，此類的模型肯定會相較 general 的模型表現在此主題領域上更好，類似的模型包括(BERTweet、RoBERTa 等)，不論是使用 Tweeter 文本做 fine-tune 或是針對情緒分析做過調整，都是非常是用在此競賽中的，這個部份是我在過程中忽略的重要部分。

最後排名結果: