社群媒体探勘作業2

108521091 電執所碩一 鄭少鈞

1. How to Reproduce

① pip install -r requirements

② python train.py

③ Vram 最多 > 20gb 否則調整 Batch Size

2. Framework.

① Data Preprocessing ⇒ ① 清理 text 是空白

② author 的 Feature 不採用

③ 將 Train 分成 0.8 Train 0.2 Validation

② Modeling Concept

採用的是 BERT Pretrained Model Fine-Tune

① 將句子輸入進 BERT 中得到 768×1 的 Vector

② 並將得到的向量接到一個 DNN 做 Binary Classifier

3. 結果探討

① 中間除了 BERT，有試過 albert, Roberta，發現 Roberta 效果

② Data Preprocess 有試過將一些特殊符號過濾，甚至是 I'm = I am

You've = You are

還有 Emoji 的部分，但 Train 出的效果都不是很好，有可能是某些符號如上下引號隱含了句子的訊息，亦或是全部句子沒有完整被清除，導致標點符號被模型誤認為特徵