Programming Assignment 3

Sentiment Analysis

機械四 b07502057 李婷穎

1. 處理資料

從給定的 train.csv 檔觀察資料分布情形，發現 1:0:–1 的比例大約是 14:6:10，故推測未來預測之結果可能大多為 1 與 –1。

接著將文章內容進行文字處理：將所有文字換成小寫英文符號、去除數字、去除標點符號、使用 nltk 套件之 PorterStemmer 進行 lemmatization、去除文章中含有 nltk 套件之 stopwords 之字詞。

因為後來發現 stopwords 中有 not, no 負面字詞，若刪除可能會造成無法判別負面意義，因此我將此二字從 stopwords 中刪除，以保留在 wordvector 中。

再來將文字向量化，使用 TfidfVectorizer 套件來建立 feature array，取 max_features=2500，得到一 31860 rows × 2500 columns 之 array。

2. 建立模型

將 train.csv 用 sklearn 套件中之 train_test_split 來把整份資料分成訓練資料及測試資料兩部分進行測試，訓練及測試資料比例為 0.8:0.2。

“利用 sklearn 來建立預測模型。我使用了 RandomForest, GradientBoosting, naive_bayes 三種模型，分別得到 71.95%, 73.02%, 72.30% 的準確率，若綜合三種模型得到的準確率為 72.75%，GradientBoosting 所得出之準確率最高，因此選用 GradientBoosting 來預測 test.csv。“

因為使用 GradientBoosting 測試結果準確率僅有 40%，因此我認為此處可能是因為原先選擇的模型太複雜，在這種情況下不適用此模型。後來翻閱了建錦老師的上課講義後（Neural Network, Deep Learning）決定採用 logistic regression 來進行訓練以及預測。發現此模型訓練 training data 的準確率有到 69.2%，F1

值 51.35%。

3. 預測

將給定的 test.csv 檔按照處理 train.csv 之步驟，去除多餘文字並向量化、建立特徵，後將結果放入上步驟得到之模型中得出預測結果。