學號：R06942077 系級： 電信碩一 姓名：洪健鈞

1. (1%) 請說明你實作的 RNN model, 其模型架構、訓練過程和準確率為何？(Collaborators:)

答： Training label data 的部分用 20 個 epcoh, 有設定 early stop 的機制來防止 overfitting, moniter 為 val_acc, optimizer 為 rmsprop, loss function 為 binary_crossentronpy。準確率為 0.82 左右。

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 40)                0
_________________________________________________________________
embedding_1 (Embedding)      (None, 40, 256)           5120000
_________________________________________________________________
lstm_1 (LSTM)                (None, 512)               1574912
_________________________________________________________________
dense_1 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 257
=================================================================
Total params: 6,826,497
Trainable params: 6,826,497
Non-trainable params: 0
_________________________________________________________________
```

2. (1%) 請說明你實作的 BOW model, 其模型架構、訓練過程和準確率為何？(Collaborators:)

答：Training label data 的部分用 20 個 epcoh, 有設定 early stop 的機制來防止 overfitting, moniter 為 val_acc, optimizer 為 rmsprop, loss function 為 binary_crossentronpy。準確率僅為 0.73。

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 40)                0
_________________________________________________________________
dense_1 (Dense)              (None, 2048)              83968
_________________________________________________________________
dropout_1 (Dropout)          (None, 2048)              0
_________________________________________________________________
dense_2 (Dense)              (None, 2048)              4196352
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 2049
=================================================================
Total params: 4,282,369
Trainable params: 4,282,369
Non-trainable params: 0
_________________________________________________________________
```

3. (1%) 請比較 bag of word 與 RNN 兩種不同 model 對於"today is a good day, but it is hot"與"today is hot, but it is a good day"這兩句的情緒分數，並討論造成差異的原因。
(Collaborators:)
答：

"today is a good day, but it is hot" RNN 分數為 0.284691, BOW 分數為 0.363076。

"today is hot, but it is a good day"RNN 分數為 0.969912, BOW 分數為 0.363076。

>>>>>因為 RNN 會由整句判斷情緒分數, 會有前後文意的判斷及連接詞判斷, 故會因為語序的不同而有不同的分數。 然而從 BOW, 兩筆資料看起來是一模一樣的故無法分辨, 而產生相同的情緒分數。

4. (1%) 請比較"有無"包含標點符號兩種不同 tokenize 的方式，並討論兩者對準確率的影響。
(Collaborators:)
答：

套用 RNN 的 model,

有標點符號的 public score 為 0.80441, private 為 0.80407

無標點符號的 public score 為 0.80917, private 為 0.80717

此為用過 simple baseline 跑實驗的結果, 看下來無標點符號的 tokenize 準確率稍為高一些, 或許無標點符號會去除一些雜訊, 但兩者差距不是很大。

5. (1%) 請描述在你的 semi-supervised 方法是如何標記 label，並比較有無 semi-surpervised training 對準確率的影響。
(Collaborators:)
答：

我的 semi-supervised learning 使用 threshold 來標記 label, 我的

Threshold 設定為 0.17, predict 出來小於 0.17 或是大於 0.83 才會當成新的 label data 繼續 train。

我的 semi-supervised 每個 iteration 有 3 個 epoch,共計 10 次 iteration, 準確率可以通過 strong baseline。

而沒用 semi-supervised 的情況下,準確率只在 simple baseline 上下,可見 semi-supervised 的情況下可以為 model 製造更多的 data。前提是 threshold 設定得宜。