學號：B03902013 系級： 資工四 姓名：吳克駿

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: 助教 code

答：

模型架構：

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 64)                82176
_________________________________________________________________
activation_1 (Activation)    (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 16)                1040
_________________________________________________________________
dropout_1 (Dropout)          (None, 16)                0
_________________________________________________________________
activation_2 (Activation)    (None, 16)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 17
_________________________________________________________________
activation_3 (Activation)    (None, 1)                 0
=================================================================
Total params: 83,233
Trainable params: 83,233
Non-trainable params: 0
```

訓練過程：

先將所有的 training data，包含 label 及 nolabel，將每個詞彙透過 gensim 訓練成 256 維的 word embedding vector

在 preprocess 時設定每一行最多 25 個字，空白補 0，確保每一個句子都是 25*256 維的向量

過一層 LSTM，兩層 DNN，使用 binary_crossentropy 計算損失函數，優化器使用 adam

取 190000 個 training data，10000 個 validation data

Epoch=50,batch=128,dropout=0.4，會在中間紀錄 validation 成績最好的 model

準確率：

Training Accuracy	Validation Accuracy	Public Score	Private Score
0.8538	0.8202	0.81860	0.81662

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: Self

答：

模型架構：

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 512)               1049600
_________________________________________________________________
activation_1 (Activation)    (None, 512)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 513
_________________________________________________________________
activation_2 (Activation)    (None, 1)                 0
=================================================================
Total params: 1,050,113
Trainable params: 1,050,113
Non-trainable params: 0
```

訓練過程：

由於訓練量極大，且詞彙過多，因此我只選取出現次數最多的 2048 個詞彙做 BOW 的 index，而由於矩陣並無排列關係，因此沒有選用 CNN，而是直接兩層 DNN，由 2048 維降為 1 維。

使用 binary_crossentropy 計算損失函數，優化器使用 adam

取 190000 個 training data，10000 個 validation data

Epoch=50,batch=128,dropout=0.4，會在中間紀錄 validation 成績最好的 model

準確率：

Training Accuracy	Validation Accuracy	Public Score	Private Score
0.9709	0.7897	0.77303	0.77303

也因為字詞互相沒有關係，單純是以出現的多寡當作是依據，因此效果並不好

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: Self

答：

	today is a good…..	Today is hot…..
RNN	0.9514	0.8448
BOW	0.3357	0.3357

由於 RNN 所使用的 word embedding vector 有詞語次序的關係，因此比較能正確判斷出語氣，做出來的結果也顯示，雖然分數不太一致，有可能是因為第一句話一開始就說今天是好天氣，但第二句先小小抱怨了一下，但 predict 出來仍是正確的。

但在 BOW 裡面，兩句話內用字完全一模一樣，因此兩句話分數完全一致，且 predict 出來的結果並不正確，推斷的原因是在 training 當中，hot 跟 but 帶給負面的效果太重，因此 model 會往壞的方面去 predict。

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:B03902101 楊力權

答：

	Training Accuracy	Validation Accuracy	Public Score	Private Score
有標點符號	0.8538	0.8202	0.81860	0.81662
無標點符號	0.8238	0.8207	0.81607	0.81535

無標點符號對於預測的結果稍差了一些，雖差距不大，但應可推斷標點符號對語句的關係及語意仍

有些許的影響。

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:B03902093 張庭維

答：

從原有 training data 訓練出一個最好的 model 後，將 no-label 的 data 來 test，但因檔案大小關係只用了其中的十萬個。若做出來的值>0.9，或者是<0.1，代表這句話夠有代表性，因此我將這句話加入 training data 中，之後再重新訓練一次。

結果：

Training Accuracy	Validation Accuracy	Public Score	Private Score
0.8722	0.8197	0.81873	0.81725

由於新加入的 label 都是由原本的 model 所生出的，因此所做出來的 training accuracy 確實變高了，但我選擇的 valid data 是原有已標註好的 label，做出來的結果僅有微幅上升，可能的原因是利用的 no label data 太少，導致還是很像原本做出來的結果。