學號：B03902093 系級： 資工四 姓名：張庭維

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:)

Model Structure:

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 50, 32)            10298336
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               53200
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 101
=================================================================
Total params: 10,351,637
Trainable params: 10,351,637
Non-trainable params: 0
_________________________________________________________________
None
```

Loss	Optimizer	Epoch	Batch size	Public	Private
BinaryCrossEntropy	adam	2	64	0.80822	0.80609

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:)

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 256)               5595136
_________________________________________________________________
dense_2 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 129
=================================================================
Total params: 5,628,161
Trainable params: 5,628,161
Non-trainable params: 0
_________________________________________________________________
None
```

選擇出現次數多於 5 次的作 Bag of Word，再將它送進上列 model 中 train，得下表結果

Loss	Optimizer	Batch size	Public	Private
BinaryCrossEntropy	adam	64	0.78862	0.78414

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:)

	RNN	BOW
today is a good day…	0.21715	0.70437
today is hot…	0.84661	0.74615

發現，對於 Bag of word 而言，兩個句子會形成一模一樣的 vector，送進 model 裡照理會 predict 出一模一樣的結果，但可能因為標點的位置不同，所以結果略有差異，但大致來說沒有太大的差別。可看出 RNN 對於有序的語言詞句來說，會得到較好的結果。

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:)

將主要的標點去掉後，用第一題的架構，再 train 一次，得到下列結果，可看出有標點略佳，可能是因為標點在語意中，具有意義。

	public	private
有標點	0.80822	0.80609
無標點	0.80161	0.79947

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:)

用第一題裡做出的 model 對 no label 的 training data 做 predict，並將 sigmoid 後得出來的值，用不同的 threshold train 出不同 model

threshold	val_acc	public	private
0.8	0.8342	0.81207	0.80911
0.9	0.8411	0.81166	0.81035

當 threshold 是 0.8 時，多出了約 70 萬筆 semi supervised data，因為多的 data 都是較符合原本 model 的，所以從原本 label 的 data 切出的 validation data accuracy 會偏高，送上 kaggle 之後則 accuracy 略為上升。Threshold 是 0.9 時，多出了約 50 萬筆 semi supervised data，從原本 label 的 data 切出的 validation data accuracy 會較 0.8 時更高，可能是因為多出來的 data 更符合原本的 model，但上傳 public 結果卻略差，可能略有 overfit 原本 label data 的可能。