學號：B03901149 系級： 電機四 姓名：陳咸嘉

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: 無)

答：

模型架構：本次作業使用 Word Embedding 作為主要架構。詳細架構如圖，Embedding Layer 採用 50 維的 word vector，training data 和 testing data 中出現次數小於 50 次的 word 則不被訓練。Padding 的大小為 30，即長度超過或不足 30 者都會補足到 30。LSTM 為雙向 100 個 units，Dropout 比例為 0.3。LSTM 後接上 DNN 來做訓練，activation function 為 ReLU。Loss 為 binary crossentropy，optimizer 為 adam。

訓練過程：

經過多次實驗發現，超過 30 個 epoch 容易會 overfitting，因此把 epoch 數量設在 30。訓練過程中採用 semi-supervise 的方法(如第五題)。

準確率：80.1% (Validation set)，80.3% (Kaggle)。

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 30, 50)	660350
lstm_1 (LSTM)	(None, 30, 100)	60400
lstm_2 (LSTM)	(None, 100)	80400
dense_1 (Dense)	(None, 256)	25856
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 64)	16448
dense_3 (Dense)	(None, 16)	1040
dense_4 (Dense)	(None, 1)	17

Total params: 844,511
Trainable params: 844,511
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: 無)

答：

模型架構：詳細架構如圖，因記憶體大小問題， training data 和 testing data 中出現次數排名前 500 的 word 才會被納入字典。轉換成每筆資料 500 維向量後，接上 DNN 來做訓練，activation function 為 ReLU。Loss 為 binary crossentropy，optimizer 為 adam。

訓練過程：

這種方法更容易 overfitting，所以把 epoch 數量設在 10。訓練過程中採用 semi-supervise 的方法(如第五題)。

準確率： 75.3% (Validation set)。

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 256)               128256
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_2 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_3 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_4 (Dense)              (None, 32)                2080
_________________________________________________________________
dense_5 (Dense)              (None, 16)                528
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 17
=================================================================
Total params: 172,033
Trainable params: 172,033
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: 無)

答：以下列出兩句話在兩種模型下的機率值

	Bag of word	RNN(Embedding)
good day but hot	0.7990	0.3657
hot but good day	0.7990	0.9510

因為 Bag of word 僅統計每個字的出現次數，所以兩句話轉成的 vector 一樣，預測機率值自然也會相同。RNN 方法，因為考慮前後文不同單字出現的關聯性，所以兩句話預測出來的機率值換成 label 剛好相反。

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: 無)

答：

針對第一題 RNN 模型進行訓練，僅改變有無分析標點符號。

	無分析標點(同第一題)	有分析標點
Validation 平均準確率	80.1%	80.9%
Kaggle 準確率	80.3%	80.7%

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: 無)

答：

先將有標記 label 的 training data 做 RNN 的訓練之後(第一題的模型)，對沒有 label 的 training data 做預測。每筆預測結果都有一個機率值，當此機率值低於 threshold 或高於 (1-threshold)時，會將該筆 data 給定標籤 0 或 1。如，threshold 為 0.2 時，機率高於 0.8 的標為 1，低於 0.2 的標為 0，並將這些 data 串接到原本有 label 的 data 中。持續進行訓練，直到所有未標明 label 的 data 少於某個特定數量為止。

透過觀察 validation 的準確率，未加上未標明 label 的準確率約為 78~79%，加上約 80 萬筆以後，準確率依然維持在 78~79%附近，且 epoch 越多也有 overfitting 的情形發生。推測原因為，採用文本內容的字典相近，故準確率沒有太大的變化。