學號：R06922113 系級： 資工所碩一 姓名：陳宣伯

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:李岳庭、林冠廷)

答：

✓ 資料前處理

我使用 Gensim 套件進行 word2vector，參數如以下：

model = Word2Vec(sentences, size=100, window=5, min_count=1)

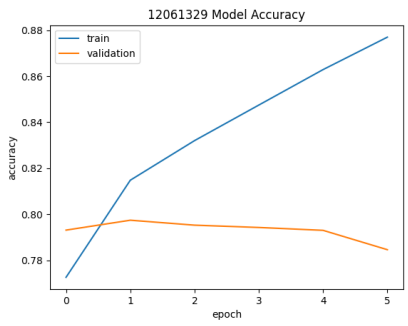
再來將所有向量 padding 到同一長度 40，有利於計算平行化

✓ 模型架構

```
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 40)                0
_________________________________________________________________
embedding_1 (Embedding)      (None, 40, 128)           2560000
_________________________________________________________________
bidirectional_1 (Bidirection (None, 40, 1024)          2625536
_________________________________________________________________
bidirectional_2 (Bidirection (None, 512)               2623488
_________________________________________________________________
dense_1 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 257
=================================================================
Total params: 7,940,609
Trainable params: 7,940,609
Non-trainable params: 0
_________________________________________________________________
```

✓ 訓練過程

辨識率：82.063%



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: 李岳庭、林冠廷）

答：

✓ 資料前處理

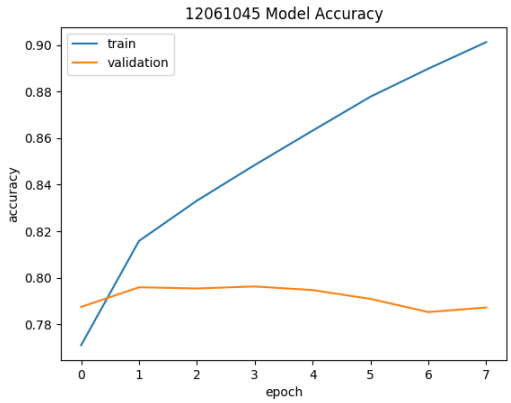
同上第 1 題的 RNN 模型討論

✓ 模型架構

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 40)                0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               10496
_________________________________________________________________
dense_2 (Dense)              (None, 256)               65792
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_2 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 129
=================================================================
Total params: 109,313
Trainable params: 109,313
Non-trainable params: 0
```

✓ 訓練過程

辨識率：75.12%



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: 李岳庭、林冠廷)

✓ 下表是兩個模型對以上兩句話的輸出

semtence	BOW	RNN
Today is a good day, but it is hot	94.67%	29.66%
Today is hot, but it is a good day	94.67%	98.43%

✓ 討論原因

由於 BOW 只考慮一句話裡面出現的單詞，而不考慮單詞間的相對位置，也就是不可慮上下文，所以以上兩句話通過 BOW 模型結果都是一樣的。但是 RNN 模型會記憶上下文，所以兩句話會有截然不同的結果

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: 李岳庭、林冠廷)

✓ 結果如下

Filter	Accuracy
'!"#$%&()*+,-./:;<=>?@[\\]^_`{|}~\t\n'	77.69%
'\n'	81.31%

✓ 討論結果

去除標點符號有較好的辨識率，我認爲應該當考慮某些標點符號時，一句話的語氣可以更加明顯，例如'!'，就可能有驚嘆、驚訝的感覺

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: 李岳庭、林冠廷）

✓ 實踐 semi-supervised 方法

直接將 no label 的 training data 丟進之前用有標籤的 20 萬筆資料建的模型，如果預測值大於 0.8，才會被視為 1，預測值小於 0.8，會被視為 0，最後把這些 130 萬新貼的資料與原本 20 萬資料一起再 train 一個新的模型

✓ 結果如下表

Semi-supervised	Accuracy
True	79.97%
False	80.53%

✓ 結果討論

有做 semi-supervised 的結果表現不如原本的模型，我認為這應該是 self-train 的缺點之一，我認為原因是本質上機器還是在學習那 20 萬筆資料，且這個模型辨識率就是 80.53%，這麼一來把這個模型拿去預測新的資料辨識率都不會穩定的成長，會一直在 80%附近徘徊