

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: b04901060 黃文聰)

答：word2vec維度：256

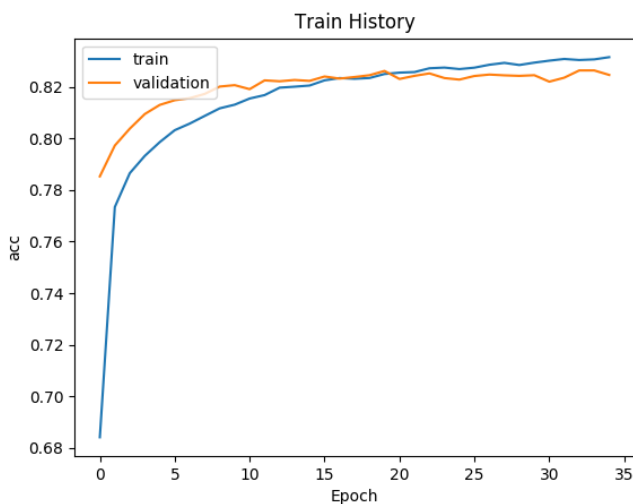
Maximum number of words：25

【模型架構】

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 25, 256)	0
gru_1 (GRU)	(None, 25, 512)	1181184
gru_2 (GRU)	(None, 512)	1574400
dense_1 (Dense)	(None, 512)	262656
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131328
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 1)	257
Total params: 3,149,825		
Trainable params: 3,149,825		
Non-trainable params: 0		
Train on 180000 samples, validate on 20000 samples		

試過多層LSTM和GRU後發現，兩層GRU三層Dense的結果最好。

【訓練過程】



【準確度】

	Accuracy
Private	0.82369
Public	0.82403
Average	0.82386

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: b04901060 黃文聰)

答：Maximum number of words：25

【模型架構】

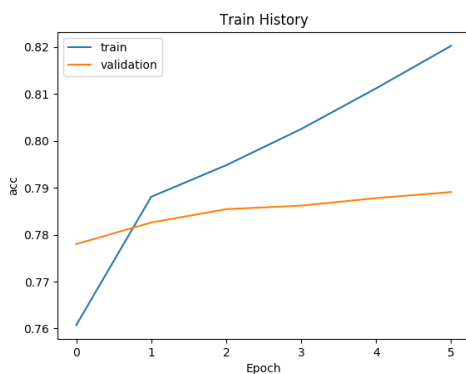
```
compile model...  


| Layer (type)         | Output Shape | Param # |
|----------------------|--------------|---------|
| input_1 (InputLayer) | (None, 3500) | 0       |
| dense_1 (Dense)      | (None, 64)   | 224064  |
| dense_2 (Dense)      | (None, 1)    | 65      |

  
Total params: 224,129  
Trainable params: 224,129  
Non-trainable params: 0  
  
Train on 180000 samples, validate on 20000 samples
```

model由dense_1(64)和dense_2(1, activation='sigmoid')組成。

【訓練過程】



【準確度】

	Accuracy
Private	0.79238
Public	0.79213
Average	0.792255

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: None)

答：line_1："today is a good day, but it is hot"

line_2："today is hot, but it is a good day"

	RNN	BOW
line_1	0.1105805	0.6282089
line_2	0.9462007	0.6193492

【RNN】GRU會考慮字詞前後關係，情緒分數顯示，第一句偏向負面情緒，而第二句偏向正面情緒。就第一句而言，可能是因為good在but之前，而機器有learn到正向詞彙後面如果接but，代表事情朝負面發展，因此會給出偏向負評的分數。

【BOW】BOW只會數單字出現次數，與前後順序無關，因此兩句的BOW表達式相同，情緒分數應該要一樣，但我predict出來的結果有些許差距，差了約0.008，可能是電腦運算上的小差距。在我的BOW model中得到的是偏向正面情緒。

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。

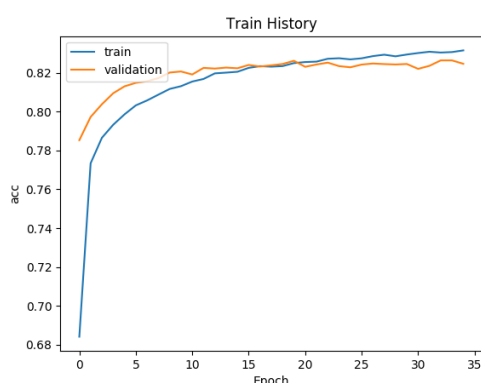
(Collaborators: None)

答：word2vec維度：256

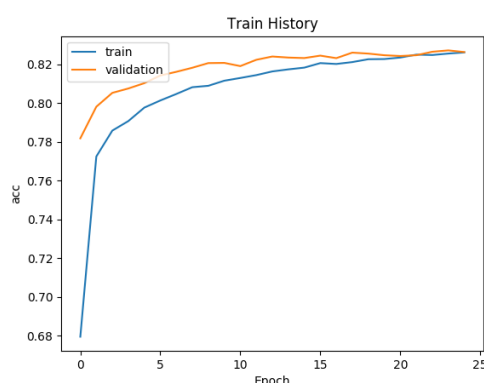
	有標點符號	無標點符號
Private	0.82369	0.82268
Public	0.82403	0.82375
Average	0.82386	0.823215

Maximum number of words：25

【訓練過程】 有標點符號



無標點符號



實驗顯示，有標點符號的準確率稍高0.06%，可能因為twitter貼文比較日常，網友常以標

點符號表達情緒，因此加入標點符號去判斷能增加準確率。

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

(Collaborators: None)

答：

threshold : 0.9

【標記方法】

一開始先以有label的data去train，並將train出來的model對nolabel data做predict，並將predict結果 >0.9 或 <0.1 的data加入原先有label的training data，一起train 20個epoch，再predict。上述步驟總共執行7輪。

	無semi	有semi
Private	0.82369	0.81101
Public	0.82403	0.81171
Average	0.82386	0.81136

實驗結果顯示，semi supervised的結果比沒有semi supervised還差。