HW5

學號：R06922086 系級： 資工所 姓名：林凡煒

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？

Model Structure

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 40)	0
embedding_1 (Embedding)	(None, 40, 128)	2560128
lstm_1 (LSTM)	(None, 40, 512)	1312768
lstm_2 (LSTM)	(None, 40, 512)	2099200
lstm_3 (LSTM)	(None, 512)	2099200
dense_1 (Dense)	(None, 256)	131328
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 1)	257

Training Process

Preprocessing: 先將 input data 去掉標點符號後，轉換成長度為 40 的 sequence vector，在透過 Word2Vec 變成 word vector。然後再丟入 RNN Model 中。

Parameters:

epochs = 20, batch size = 256,

model check point: monitor=’val_acc’, save_best_only=True

early stopping: monitor=’val_acc’, patience=3

validation data: 1 / 10 training data

Accuracy

Kaggle’s private / public score: 0.82158 / 0.82278

2. *請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？*

Model Structure

Layer (type)	Output Shape	Param #
dense_8 (Dense)	(None, 1024)	2049024
activation_4 (Activation)	(None, 1024)	0
dropout_4 (Dropout)	(None, 1024)	0
dense_9 (Dense)	(None, 512)	524800
activation_5 (Activation)	(None, 512)	0
dropout_5 (Dropout)	(None, 512)	0
dense_10 (Dense)	(None, 256)	131328
activation_6 (Activation)	(None, 256)	0
dense_11 (Dense)	(None, 1)	257

Training Process

Preprocessing: 去掉標點符號後，轉成 2000 維的 BOW vector，再丟入 DNN Model。

Parameters:

epochs = 20, batch size = 256,

model check point: monitor=’val_acc’, save_best_only=True

early stopping: monitor=’val_acc’, patience=3

validation data: 1 / 10 training data

Accuracy

Kaggle’s public/private score: 0.79020 / 0.79081

3. *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。*

	Today is a good day, but it is hot	Today is hot, but it is a good day
Bag of word	0.6196314	0.6196314
RNN	0.7671631	0.9555416

BOW model 下對兩個句子的預測結果會是一樣的，但 RNN 對後者的預測比起前者，正情緒的傾向是比較明顯的。
會造成這種差異，是因為 RNN model 有記憶的特性，會考慮句子中詞彙的順序。

4. *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。*

	Kaggle’s public score	Kaggle’s private score
有標點符號	0.81733	0.81931
無標點符號	0.80025	0.79625

由結果可以看出，有標點符號的結果會是比較好的。
猜測是基於有些標點符號可能是會影響語意的，比如說質問或者反諷的句子通常後面會接上問號，此時若有標點符號可能可以幫助 model 對情緒的判斷。

5. *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。*

	Kaggle’s public score	Kaggle’s private score
Semi-supervised	0.82278	0.82158
Non-semi-supervised	0.82006	0.82107

首先將 non-labeled data 切成 10 份，編號 1~10。
取出編號 1 的 data 作 prediction，若是結果 > 0.8 or < 0.2 ，則將 data 加入 training data 中進行 retrain。
接下來依編號序號 1~ 10 一次取兩份 non-labeled data 進行 prediction。這時編號 1 的 data 會有兩次 prediction 的結果，取出 結果 > 0.8 or < 0.2 的 data，觀察是否與前一次相同，若是相同則加入 training data 中作 retrain。
由結果可以看出，semi-supervised 的 performance 比較好。