學號：B04901117 系級： 電機三 姓名：毛弘仁

(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：

Layer (type)	Output Shape	Param #
embedding_11 (Embedding)	(None, 500, 32)	160000
lstm_11 (LSTM)	(None, 100)	53200
dense_11 (Dense)	(None, 1)	101

總共 213,301 parameters，我下一題的 BOW model 根據這個數量做設計。Word dictionary 只保留 unlabeled data 中出現次數最高的 5000 字，並且將標點符號移除。

Training accuracy = 0.808 、validation accuracy = 0.801、Kaggle test accuracy = 0.8026，其中 validation set 是 training set 最後 10% 的 data 。

Batch size = 64、epochs = 3、word vector dimension = 32、input sentence pad to length 39（這是 training_label、testing_data 句子最大的長度）word indices、LSTM output dimension = 100、LSTM input dropout 0.2、LSTM recurrent dropout = 0.2、optimizer = Adam。

(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：

Layer (type)	Output Shape	Param #
dense_23 (Dense)	(None, 43)	215043
dense_24 (Dense)	(None, 43)	1892
dense_25 (Dense)	(None, 1)	44

總共有 216,797 parameters，和 RNN model 差不多。

Training accuracy = 0.794、validation accuracy = 0.78。其他 hyperparameters 和 RNN model 一樣，只有 epochs 改成 10（training loss 和 validation loss 在這個時候差不多 converge 了）。

(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：

BOW model 對於兩句的預測都是 positive（output 0.66），因為沒有在乎 word order；RNN model 對第一句的預測是 negative（output 0.4，可能因為 hot 不是太負面的字所以 output 沒有更低；我換成其他明顯負面的字，output 就會降低），對第二句的預測是 positive（0.87），能夠根據 word order 做出不同判斷，因為 RNN input 會有時序性，越後面的字應該影響越大。

(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：

含標點符號的 RNN model

Training accuracy = 0.8114、validation accuracy = 0.808，比不含標點符號的 accuracy（0.808、0.801）還高一點點，在 Kaggle 上的 test accuracy 也高了一些，從 0.802 跳到 0.804。可能是 Twitter 上面的標點符號和 sentiment 有些相關性，我跑了一些分析，發現 training data 中，含有「!」的句子，有 0.6 的機率是 positive sentiment，「!」也可能有加強語氣的作用。

含標點符號的 BOW model

Training accuracy = 0.793、validation accuracy = 0.78，和不含標點符號的 accuracy 差不多。從這個結果來看，即便含有「!」的句子有 0.6 機率是 positive，但是 BOW model 沒有受太大的影響，因此如果加上「!」有 performance gain，那應該是在 RNN 中和其他字產生交互作用，而不是符號本身有影響。

(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: 無)

答：

如果 RNN model 對於 unlabeled data 的 prediction > 0.8，我把這筆 data 加進 training data，label 設成 1；如果 prediction < 0.2，label 設成 0。這樣會使得原本 200,000 筆 labeled training data 增加到 912,763 筆，大小為原本的 4.5 倍左右。

訓練結果是 training accuracy = 0.95、Kaggle 上的 test accuracy = 0.803，可見雖然 training accuracy 增加許多，在 test data 上卻沒有顯著的 accuracy gain 。或許是因為我用的 semi-supervised 方法過於強化既有的 label，沒有對 model 產生很大的變化。