學號：R06725005 系級： 資管碩一 姓名：郝思喬

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:)

   答：

   先使用 gensim 的 word2vec 將 label 與 nonlabel 的文字轉成 200 維的向量，將這些值當成 Embedding Layer 的 embedding matrix，再將 label data 經過 tokenizer 當成 input data。模型架構是先在第一層接上自己用 gensim 得到的 embedding，並 trainable 設為 False，接著接三層 LSTM，每層的 dropout、recurrent_dropout 都設為 0.2、units=128，再接上一層 Dense、units=128、使用 relu，最後一層為 units=1 的 Dense，使用 sigmoid 當成 output。

   訓練過程除了 tokenizer 與 gensim 的前處理外，另外也要先把 tokenizer 存下來以便在 testing 的時候 tokenizer 的對應不會出錯。另外還做了 10-fold cross validation。

   單一 model 準確率為 0.82784，10-fold model 準確率為 0.83565。

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 37, 200)	50126200
lstm_1 (LSTM)	(None, 37, 128)	168448
lstm_2 (LSTM)	(None, 37, 128)	131584
lstm_3 (LSTM)	(None, 128)	131584
dense_1 (Dense)	(None, 128)	16512
dense_2 (Dense)	(None, 1)	129

Total params: 50,574,457
Trainable params: 448,257
Non-trainable params: 50,126,200

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:)

   答：

   BOW model 是 training set 使用 Tokenizer 先 fit 過後，經過 pad_to_sequences 以及 sequences_to_matrix 過後，再丟入模型中。模型架構為一層 256 的 Dense，Dropout 為 0.3，在接上一層 1 的 Dense 當成 output，準確率為 0.7815。

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 37, 200)           50126200
_________________________________________________________________
lstm_1 (LSTM)                (None, 37, 128)           168448
_________________________________________________________________
lstm_2 (LSTM)                (None, 37, 128)           131584
_________________________________________________________________
lstm_3 (LSTM)                (None, 128)               131584
_________________________________________________________________
dense_1 (Dense)              (None, 128)               16512
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 129
=================================================================
Total params: 50,574,457
Trainable params: 448,257
Non-trainable params: 50,126,200
_________________________________________________________________
```

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:)

   答：

	BOW	RNN
"today is a good day, but it is hot"	0.72257692	0.78329289
"today is hot, but it is a good day"	0.72257692	0.68493307

   BOW 因為字詞的順序不管怎麼對換會對應到相同的 input，因此兩句會得到相同的結果。但是 RNN 會因為字詞先後順序有差別，因此會得到不同的結果。在此例子中因為 it is a good day 對於正面情緒的作用較大，因此預測出來皆為正向。

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:)

   答：

   tokenize 濾掉標點符號得出來的準確率為 0.79831，沒有濾掉標點符號得出來的準確率為 0.80357。沒有濾掉標點符號準確率較高，可能是因為這些標點符號也會代表一些特徵，例如驚嘆號出現較有可能代表這句話越正向。

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:)

   答：

   在 semi-supervised 的方法中，在訓練時每個 epoch 結束後都先將 nolabel 的 data 預測一次，將其 output 為 0.95 以上的標記其 label 為 1，output 為 0.05 以下的標記其 label 為 0，再將這些有被標為 0 或 1 的 data 加入 training set 一起再下去訓練。沒有使用 semi-supervised 的 model 準確率為 0.79831，使用後為 0.80357，有進步但是無明顯差別。