學號：R05631027 系級：生機碩二 姓名：楊皓文

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：

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 40, 256)	65749504
bidirectional_1 (Bidirection	(None, 40, 512)	1050624
bidirectional_2 (Bidirection	(None, 512)	1574912
dense_1 (Dense)	(None, 256)	131328
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 256)	65792
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 1)	257

Total params: 68,572,417
Trainable params: 2,822,913
Non-trainable params: 65,749,504

前處理: 使用 genism 中的 Word2Vec，將 training、testing、nolabel 的單字轉換為 256 維的向量，Tokenizer 使用的 filters='\n'(有表點符號)，所有句子都 padding 到 40 的長度，全部的字詞都有採用。

RNN model training: 1 層利用 Word2Vec 產生的 embedding layer，2 層 bidirectional LSTM layer，3 層 dense layer(如上圖)。30 個 epoch，optimizer 為 adam， loss function 為 binary_crossentropy。

Validation accuracy: 0.83725

Kaggle private score: 0.83220

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：

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 1024)	5121024
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 512)	524800
dropout_2 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 256)	131328
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 1)	257

Total params: 5,777,409
Trainable params: 5,777,409
Non-trainable params: 0

前處理: Tokenizer 使用的 filters='!"#$%&()*+,-./:;<=>?@[\\]^_`{|}~\t\n' (無表點符號)，只採用出現最多的前 5000 個字詞。

BOW model training: 4 層 dense layer(如上圖)，20 個 epoch，optimizer 為 adam，loss function 為 binary_crossentropy。

Validation accuracy: 0.79550

Kaggle private score: 0.79543

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：

	"today is a good day, but it is hot"	"today is hot, but it is a good day"
RNN val_acc	0.39115831	0.98177034
Bag of word val_acc	0.66959184	0.66959184

輸出的準確率如上，可以看出 RNN 由於經過有前後關係的 LSTM layers，因此文次排序不同輸出的結果差異很大，而也較接近原本語意的情緒。而 Bag of word 因為是計算單字出現的次數，因此兩者的分數會完全一樣，無法判別。

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：

設定 Tokenizer 的 filters 分別為，filters='!"#$%&()*+,-./:;<=>?@[\\]^_`{|}~\t\n'和 filters='\n'，分別是一個沒有標點符號，另一個有標點符號(只有去掉換行符號)。

結果為:沒有標點符號的 Tokenizer kaggle private score 為 0.83220，有標點符號的 Tokenizer kaggle private score 為 0.82522。

從結果可以看出有包含標點符號的 Tokenizer 會有較好的表現，推測是一些情緒的標點符號，如:!、?...等，在情緒判別上是有用特徵。

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：

利用表現最好的 RNN model 先產生預測值，預測值大於 0.95 標記為 1，小於 0.05 標記為 0，其他則不採用。總共從 1178614 中取出 581860 筆 semi-supervised 的資料，在加入 training 的資料集，最後再利用相同的 RNN model 從頭訓練。

最後訓練出來的成果: kaggle private score 為 0.82972，略差於原本 training 的資料集 (kaggle private score 為 0.83220)。

可能由於 nolabel 的資料集較多，使用 RNN model 預測時反而是增加其中的 bias，使得 semi-supervised 的資料最後並沒有達到加分的效果。