學號：R06922075 系級： 資工碩一 姓名：翁瑋

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 #
gru_1 (GRU)	(None, 40, 64)	37056
gru_2 (GRU)	(None, 40, 32)	9312
gru_3 (GRU)	(None, 32)	6240
dense_1 (Dense)	(None, 128)	4224
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8256
dropout_2 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 2)	130

Total params: 65,218
Trainable params: 65,218
Non-trainable params: 0

圖片是我的模型架構，使用了3層GRU(0.15的dropout跟0.15的recurrent_dropout)，後接3層DNN，每層都加0.25的dropout，activation使用relu，最後一層用softmax，train16個epoch，validation0.05，最後在local的validation accuracy為0.8223，kaggle上為public : 0.8246

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：使用keras內建的tokenizer去做，選取出現頻率最高的2000個字去做編碼(選太多記憶體會不足)，fit完data後，再用texts_to_matrix去將每句轉成相對應的BOW vector，後接4層的DNN，沒有任何dropout，最後在local的validation accuracy為0.7720，kaggle上為public : 0.77155 private : 0.77251

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：
Sentence1 = "today is a good day, but it is hot"
Sentence2 = "today is hot, but it is a good day"

	BOW model(77.15%)	RNN model(82.46%)
Sentence1	0.6342419	0.46814963
Sentence2	0.6342419	0.88462764

對於BOW model來說，這兩句話的字出現的數量都相同，所以在編碼的時候彙編成一樣的vector，於是predict出來的機率也是一樣，而RNN model我這裡採用的是gensim的編碼，因為每個word的向量編碼有考慮前後字的關係，而在RNN training的部分也有考慮前後字的關係，所以可以看到兩個predict出來的機率並不相同，甚至於最後label上的情緒也是相反的，對於我來說，也認為RNN的predict比較有道理

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：

再結果上看來，沒有標點符號的準確率較低，低了大約0.7%，對於這個結果，我猜測是因為標點符號對於句子的情緒也是有影響的，雖然沒有絕對，但通常不同的標點符號在不同情緒的句子中是有一定規律可循的

	public	private
有標點符號	0.82460	0.82407
沒有標點符號	0.81738	0.81792

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：在使用原本的架構做training後，對100萬筆的data做預測，在出來的機率高於0.85的情況下進行標記，得到約55萬筆的label data，再加上原本的20萬筆data去做training，在相同架構下，約可以增加0.5%的準確率

	public	private
普通RNN	0.80134	0.79809
semi-supervised	0.80772	0.80368