學號：B03901098 系級： 電機四 姓名：王建翔

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: Sample Code)

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 40)                0
_________________________________________________________________
embedding_1 (Embedding)      (None, 40, 256)           5120000
_________________________________________________________________
gru_1 (GRU)                  (None, 40, 512)           1181184
_________________________________________________________________
gru_2 (GRU)                  (None, 512)               1574400
_________________________________________________________________
batch_normalization_1 (Batch (None, 512)               2048
_________________________________________________________________
dense_1 (Dense)              (None, 256)               131328
_________________________________________________________________
batch_normalization_2 (Batch (None, 256)               1024
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 257
=================================================================
Total params: 8,010,241
Trainable params: 8,008,705
Non-trainable params: 1,536
_________________________________________________________________
```

一開始先dim為40的字句向量輸入，後經過embedding layer輸出(40,256)的embedding向量，之後經過兩層的GRU，一層全連接層，前後各有一個Batchnormalization層，然後Dropout，最後通過sigmoid function得到最後結果。

訓練過程切10%的資料為validation data，loss funciton為cross_entropy，同時也設立EarlyStopping和ReduceLR，patience分別為5和2，batchsize為512。

Kaggle	Public	Private
Acc	0.80381	0.80259

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: b03901109陳緯哲)

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 2000)              0
_________________________________________________________________
embedding_1 (Embedding)      (None, 2000, 480)         960000
_________________________________________________________________
dense_1 (Dense)              (None, 2000, 80)          38480
_________________________________________________________________
leaky_re_lu_1 (LeakyReLU)    (None, 2000, 80)          0
_________________________________________________________________
batch_normalization_1 (Batch (None, 2000, 80)          320
_________________________________________________________________
dropout_1 (Dropout)          (None, 2000, 80)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 160000)            0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 160001
=================================================================
Total params: 1,158,801
Trainable params: 1,158,641
Non-trainable params: 160
_________________________________________________________________
```

將向量輸入後接一層embedding後接一層全連階層，後面以relu, normalization, 和dropout輔助訓練，最後經過sigmoid函數得到最後答案，訓練過程切10%的資料為validation data，loss funciton為cross_entropy。

Kaggle	Public	Private
Acc	0.77376	0.77287

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。

答：

	Bag of word	RNN
情緒分數	0.6012087/ 0.6012087	0.5442213/ 0.9862688

因為RNN對於字詞的輸入順序較為敏感，所以雖然句子的組成是同些字詞，但是

因為順序的不同會影響語意，另外Bag of word則是考慮每一個句子單字出現的個數，所以會有一樣的分數，若單就這個句子來看，RNN模型應該是判斷了but後的子句才是情緒表達的重點，所以給前者較低、後者較高的分數。

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。

   答：

   在經過比較之後，發現有將標點符號做tokenize的結果會比較好，推測是因為在一些文章中，標點符號對於語氣的表達也有很大影響，像是驚嘆號、問號如果做了tokenize可以輔助RNN更好地去判斷正負語氣。

Kaggle	With tokenize	Without tokenize
public/private	0.80381/0.80259	0.79864/0.79809

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。

   答：

   於semi-supervised，對尚未label的檔案進行label，而分數超過0.8的才標示為正向，小於0.2則標記為負向的情緒，將這些資料匯於原本的訓練資料再進行訓練，validation data一樣是切10%，總共經過5次iteration，在經過semi-supervised後的結果也比較好。

Kaggle	With semi	Without semi
public/private	0.80381/0.80259	0.79798/0.79645