HW5 Report

學號：B04502031 系級： 電機二 姓名：施力維

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:)

Preprocessing：保留包點符號，將重複的字母刪掉，將 won’ t 這類的字改成 will not。

Model 如下，使用兩層 GRU 搭配兩層 Dense 來實作，其中先使用 gensim 來對 train 與 semi 的 data 做 pretrain，min count=2, dim=100，

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 100)         5707400
_________________________________________________________________
gru_1 (GRU)                  (None, None, 128)         87936
_________________________________________________________________
gru_2 (GRU)                  (None, 128)               98688
_________________________________________________________________
dense_1 (Dense)              (None, 256)               33024
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               65792
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 514
=================================================================
Total params: 5,993,354
Trainable params: 285,954
Non-trainable params: 5,707,400
_________________________________________________________________
Train on 180000 samples, validate on 20000 samples
```

Optimizer 使用 adam，訓練過程約 20 個 epoch 可以收斂，使用 earstopping，batch size 為 256，切 0.1 來作為 validation，最後的 accuracy 為(public, private)=(0.832, 0.830)。

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:)

訓練模型如下，使用 keras 的 tokenizer 轉成 bow 後，使用三層 DNN 搭配 dropout 來做 training，一樣切 0.1 當 validation，

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 256)               1536256
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               65792
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 256)               65792
_________________________________________________________________
dropout_3 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_4 (Dense)              (None, 2)                 514
=================================================================
Total params: 1,668,354
Trainable params: 1,668,354
Non-trainable params: 0
```

Optimizer 使用 adam，使用 earstopping，batch size 為 256，切 0.1 來作為 validation，最後的 accuracy 為(public, private)=(0.798, 0.798)。

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:)

RNN model 與 BOW model 分別使用上面兩題的 model，都使用 categorical crossentropy，分別判斷好的與壞的機率，結果如下：

	good	bad
RNN	0.677	0.323
BOW	0.297	0.703

可以看到兩者的預測結果是截然不同的，這邊可以明顯看出 BOW 的缺點，當句子出現轉折等語氣時便容易判斷錯誤，good 在 train 出來後肯定是比較正面的權重居多，因此當這邊出現轉折時，便會失去精確度。

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:)

這邊有標點符號使用的是第一題的 model，沒標點的 model 其他參數也都跟第一題相同，做出來的結果如下：

	Public	Private
有標點	0.832	0.830
沒標點	0.824	0.824

結果是沒標點符號較差一些，推測是因為諸如驚嘆號、問號等對於句子的語句判斷有所幫助，因此有標點符號的 model 較高一些。

5. (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:)

在 Semi-supervised 的 model 中，是將 semi 的 prediction 中 good 或 bad 大於 0.8 的 label 設為 1 或 0(使用 categorical crossentropy)，在拿去先前 train 好的 model 繼續 train，結果如下：

	Public	Private
normal	0.832	0.830
Semi-supervised	0.834	0.833

可以看到 semi-supervised 的結果比沒有還要好一些些，證實了以 0.8 作為 threshold 所獲得的 data 具有相當的可性度，有實際的成效。