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。

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

1. (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。

1. (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: )

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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。

1. (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。

1. (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。