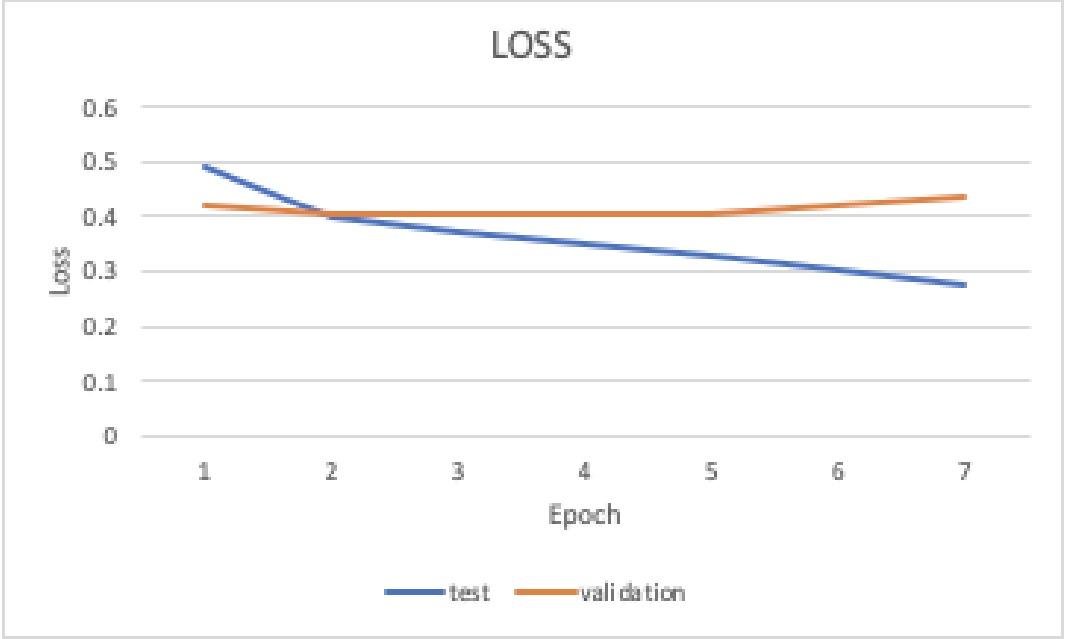
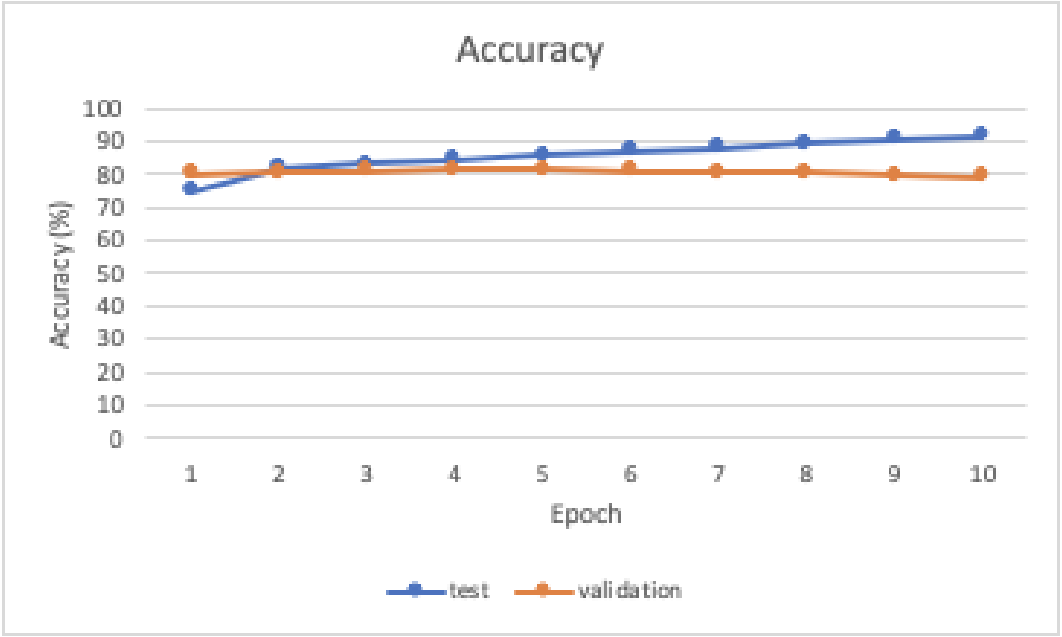
學號：b07902074　　系級：資工二　　姓名：張媄姞

1. (1%) 請說明你實作的RNN的模型架構、word embedding 方法、訓練過程 (learning curve)和準確率為何？ (盡量是過public strong baseline的model)

```
LSTM_Net(
  (embedding): Embedding(55333, 300)
  (lstm): LSTM(300, 150, batch_first=True)
  (classifier): Sequential(
    (0): Dropout(p=0.5, inplace=False)
    (1): Linear(in_features=150, out_features=1, bias=True)
    (2): Sigmoid()
  )
)
```

- 



- Embedding：
  - 用skip-gram
  - Train 一個 Word2Vec 的 model 並算出 embedding matrix
  - 藉由 embedding matrix 將輸入轉乘 model 的 input
- 準確率在validation上可以到0.8202

2. (2%) 請比較BOW+DNN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"這兩句的分數(過softmax後的數值)，並討論造成差異的原因。
   - BOW 的 model 是前面先做BOW再用下面的模型進行訓練

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1              [-1, 128, 512]         614,912
            Linear-2              [-1, 128, 128]          65,664
            Linear-3                [-1, 128, 1]             129
           Sigmoid-4                [-1, 128, 1]               0
================================================================
Total params: 680,705
Trainable params: 680,705
Non-trainable params: 0
----------------------------------------------------------------
```

"today is a good day , but it is hot" **V.S.** *"today is hot , but it is a good day"*

- RNN分別是： [0.3801] , [0.9673]
- BOW+DNN分別是： [0.6113] , [0.6113]

- 可以發現RNN對於good(正面詞)後面加上but(轉折詞)，會有較高的機率判成負面，而hot(RNN分析偏負面)後面加上but，又加上good(正面詞)則有較高的機率判成正面。
- 可以看出BOW+DNN兩句都判成正面，可能跟裡面都有good有關係，因為BOW比較不會考慮單詞之間的關聯。

3. (1%) 請敘述你如何 improve performance（preprocess、embedding、架構等等），並解釋為何這些做法可以使模型進步，並列出準確率與improve前的差異。（semi supervised的部分請在下題回答）
    - 我在preprocess的時候將n ' t 都改成了not，以及把buuusssssyyy這樣子有多個重複字的單字都改成了busy，然後本來有嘗試做stem，但是發現效果不彰，可能是因為此次的資料都較為口語化的關係。
    - 而將單字這樣處理可以讓token數減少並且保留意義，達到進步的效果。
    - 在使用preprocess之前，validation的正確率大約為0.813，而使用之後可以達到0.8201

4. (2%) 請描述你的semi-supervised方法是如何標記label，並比較有無semi-supervised training對準確率的影響並試著探討原因（因為 semi-supervise learning 在 labeled training data 數量較少時，比較能夠發揮作用，所以在實作本題時，建議把有 label 的training data從 20 萬筆減少到 2 萬筆以下，在這樣的實驗設定下，比較容易觀察到semi-supervise learning所帶來的幫助）。

- 我實作的是self-training，並將thresold設成0.7，我們可以發現準確率幾乎沒有比較高，都一樣在0.818左右，造成這樣的原因可能是因為原本的model並不是百分之百標記正確，某些錯誤判斷也因此加強了，因此即使資料變多，在準確率上也沒有表現得比較好。