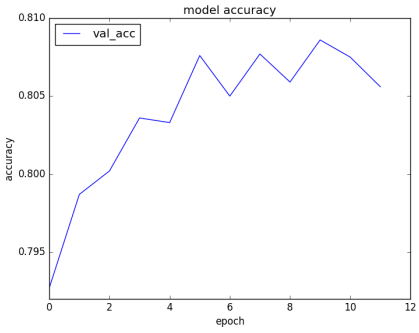
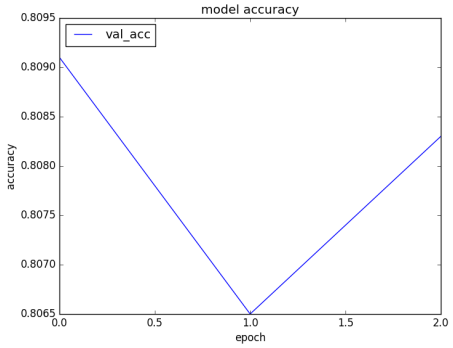
學號：R06922118 系級： 資工碩一 姓名：吳政軒

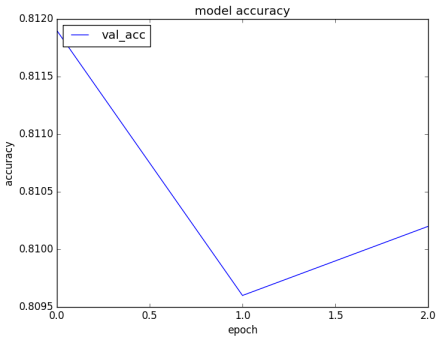
1. (1%) 請說明你實作的 RNN model, 其模型架構、訓練過程和準確率為何？(Collaborators: 黃敬庭、倪溥辰)

首先將 labeled data 經過 tokenizer 並 padding 後丟到 Embedding layer，用一層 LSTM，加上兩層 Dense，最後透過 sigmoid 得到一維的 output，train 好這個 model 後再拿去 predict unlabeled data，取機率大於 0.9 和小於 0.1 的資料來做 semi-supervised learning，semi-supervised 的部分我做了兩個 iteration，以上的 training 都有使用 early stop 和 model checkpoint 儲存最好的一次，準確率為 0.81354。

| Layer (type) | Output Shape | Param # |
|---|---|---|
| embedding_1 (Embedding) | (None, 30, 128) | 2560000 |
| dropout_1 (Dropout) | (None, 30, 128) | 0 |
| lstm_1 (LSTM) | (None, 256) | 394240 |
| dense_1 (Dense) | (None, 256) | 65792 |
| dropout_2 (Dropout) | (None, 256) | 0 |
| dense_2 (Dense) | (None, 256) | 65792 |
| dropout_3 (Dropout) | (None, 256) | 0 |
| dense_3 (Dense) | (None, 1) | 257 |



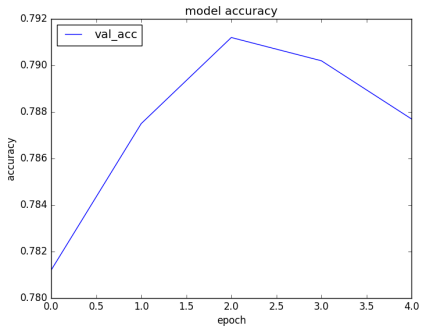



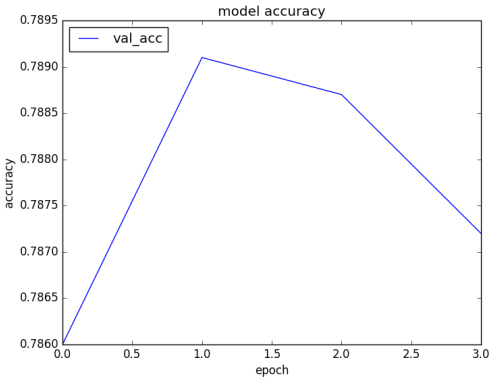



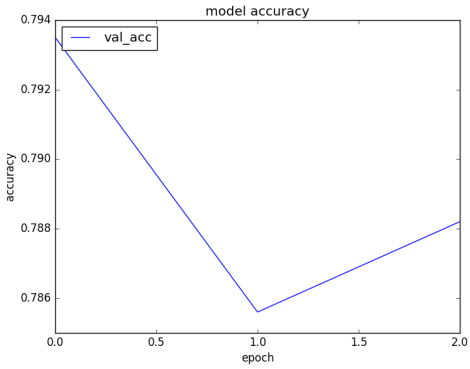

2. (1%) 請說明你實作的 BOW model, 其模型架構、訓練過程和準確率為何？(Collaborators: 黃敬庭、倪溥辰)

首先將 labeled data 經過 tokenizer，轉成 matrix 後直接丟到 Dense，其餘部分都跟第一題 RNN 一樣，準確率為 0.79037。

| Layer (type) | Output Shape | Param # |
|---|---|---|
| dense_1 (Dense) | (None, 256) | 512256 |
| dropout_1 (Dropout) | (None, 256) | 0 |
| dense_2 (Dense) | (None, 256) | 65792 |
| dropout_2 (Dropout) | (None, 256) | 0 |
| dense_3 (Dense) | (None, 1) | 257 |




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: 黃敬庭、倪溥辰)

RNN: 0.37067524, 0.9909116
BOW: 0.76914769, 0.76914769

RNN 中前者預測為負面，後者預測為正面，這是比較符合我們所預期的情緒，因為 RNN 有記憶的功能，所以比較可以分辨出情緒；而 BOW 的 model 在這兩句話預測出的分數一樣，因為這兩句話只有順序改變，句子內的字並沒有差異，也就是說 input vector 會長得一樣，也就導致這樣的結果。

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: 黃敬庭、倪溥辰)

我使用 keras 提供的 Tokenizer，調整 filters 參數來決定要不要包含標點符號，有包含標點符號的準確率為 0.81354，沒包含標點符號的準確率為 0.80747，所以標點符號還是多少能傳達一些資訊。

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: 黃敬庭、倪溥辰)

將使用 labeled data train 出來的 model 拿去 predict unlabeled data，取機率大於 0.9 和小於 0.1 的資料（有一定程度的信心）來做 semi-supervised learning，大於 0.9 就標記 1，小於 0.1 就標記 0。沒有 semi-supervised 的準確率為 0.80385，有 semi-supervised 的準確率為 0.81354。