Report

學號：B04502139 系級： 電機三 姓名：戴瑋辰

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:)

先利用所有的training data實做word2vec, 並把訓練出來的weights存起

來, 要將data丟進RNN以前, 先將每一個句子用word index表示, 並將長度補至最常的句子的長度 (zero padding) , 若出現不認識的字, 則歸類為other, 其index為-1。

所有data準備好後, 先將其丟入embedding layer並將此layer的weights設為已train好的word2vec的weights, 在接到三層RNN上, 第一、二層為LSTM (hidden size 40), 第三曾為GRU(hidden size 256), 最後將GRU的output再接到兩層hidden size為256的DNN以sigmoid作為最後一層的activation function輸出預測結果。

經過3次self-training每次10個epoch後, 準確率可達到0.82753。

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:)

將training data吃進來後，建立一個字典，每一個字都有其對應到的編號。

字典建立完成後，將所有training data 轉換成bow的形式直接丟進DNN 訓練。

此neural network共三層，每個hidden layer 的hidden size都是256，由於以bow的形式除存空間會太大，因此用keras的generator來訓練，最終在經過大約3個epoch後可以收斂到0.543%的準確率。

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:)

Accuracy：

	Today is a good day, but it is hot	Today is hot, but it is a good day
BOW	0.52158	0.52158
RNN	0.22529	0.97141

從表中可以看到使用BOW model不但train不好，也沒有辦法分辨這兩個句子的差別，主要原因是這兩個句子在BOW的模型下會長的一模一樣，因此沒有辦法區分出差異；而RNN會將上一次的結果記起來，順序不同就會有不一樣的結果，另外當用word2vec來表達一個句子時，順序不同也會是不同的句子，因此能分辨出差異。

從結果來看，可以發現RNN對於today is a good day, but it is hot判斷為負面(<0.5)，而today is hot, but it is a good day判斷為正面(>0.5)，效果非常好。

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:)

	有標點符號	無標點符號
Accuracy	0.79364	0.82753

從實驗結果可以看出再有標點符號的情況下，train出來的model表現較差；而沒有標點符號做出來比較好。推測是因為自本身的順序就能判斷句子的正負面程度，多加了標點符號把model弄的太複雜導致train不好。

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:)
semi-supervised 標記方法為下：

unlabeled data在丟進model之後，經過sigmoid出來若是大於0.8則將其歸類為class 1，若小於0.2則歸類為class 0，落在0.2 ~ 0.8之間就不理它，一樣當作是unlabeled data。

	Supervised learning	Semi-supervised learning
Accuracy	0.80037	0.82753

Semi-supervised 表現較好。