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，每一個自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 60.543%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。