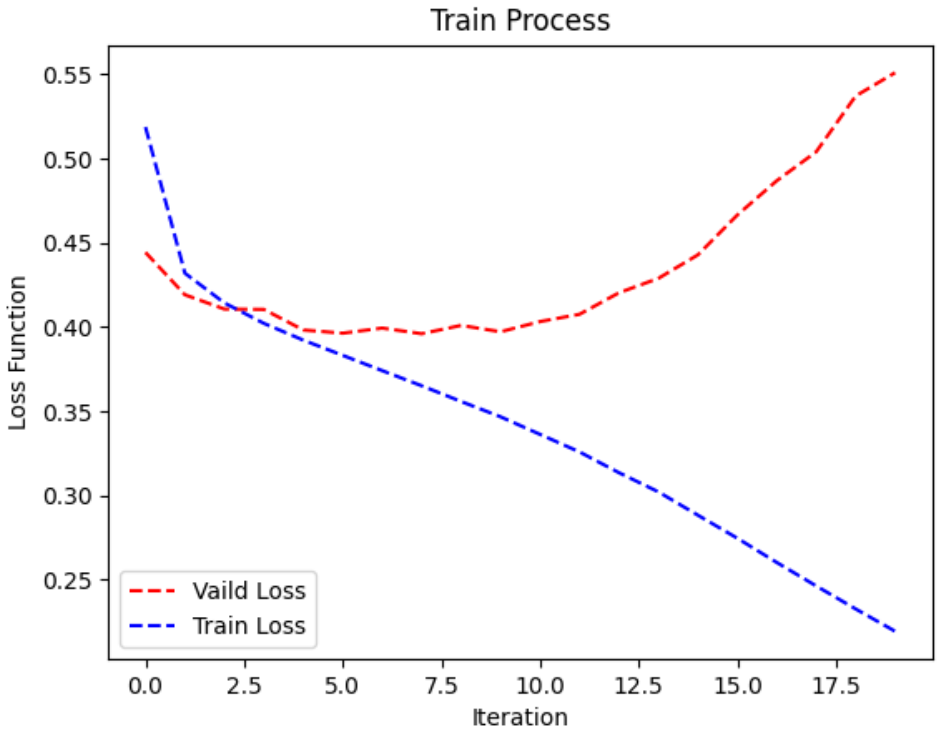
學號：B07902034　系級：資工二　　姓名：王昱凱

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)



我實作的 RNN 模型架構是與助教 colab 上的 RNN 模型相同架構，也就是一層的 LSTM model，而我也嘗試過增加層數，但準確率並沒有提升，而 word embedding 的部分我把 no label 的資料也加入一起 train，最後我將句子長度改為 30，epoch 改為 20，learning rate 改為 0.0005，在 kaggle 的準確率可以達到 0.824 左右

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。

我使用 BOW+DNN 測試後的分數皆為 0.47，認為兩句話都偏向負面，而使用 RNN 的分數則分別為 0.42 和 0.88，認為第一句話偏向負面而第二句話為正面，會造成這種差異的原因在於 BOW+DNN 只考慮了單字而沒有考慮順序和前後關

係，因此兩者才會有相同的分數，而 RNN 會考慮單字的前後順序，因此才得以判斷兩者的差異

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）

   在 prepocess 的部分我將句子長度從 20 改為 30，準確率大概能夠提升 0.01~0.015，因此我認為將句子長度設定為 30 對於判斷正負面有較好的效果，太短或太長可能都無法很準確的預測正負面，而 embedding 部分我有嘗試調整過 min_count 和 iteration 等參數，但是對於模型準確率卻沒有進步，而在模型架構上，我也嘗試過增加 LSTM 的層數，但準確率同樣也沒有進步，所以經過測試後我發現只有調整句子長度對模型進步較有效果

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）。

   在維持相同模型架構下，我先將有 label 的 data 數減少到 2 萬筆，而在 validation set 的準確率約為 73%左右，而在使用 semi-supervised 方法後，data 數增加到 110 萬筆，在 validation set 的準確率可以提升到 76%左右，而我實作的 semi-supervised 方法是 self-training，也就是先用原本的 model 對沒有 label 的 data 作預測，而分數大於 0.5 的我就將 label 設為 1，反之則設為 0，至於可以提升準確率的原因應該在於當有 label 的資料量過少時，training 得出的 model 不夠準確，因此藉由這些沒有 label 的 data 作 self-training 來彌補資料量不足的情況