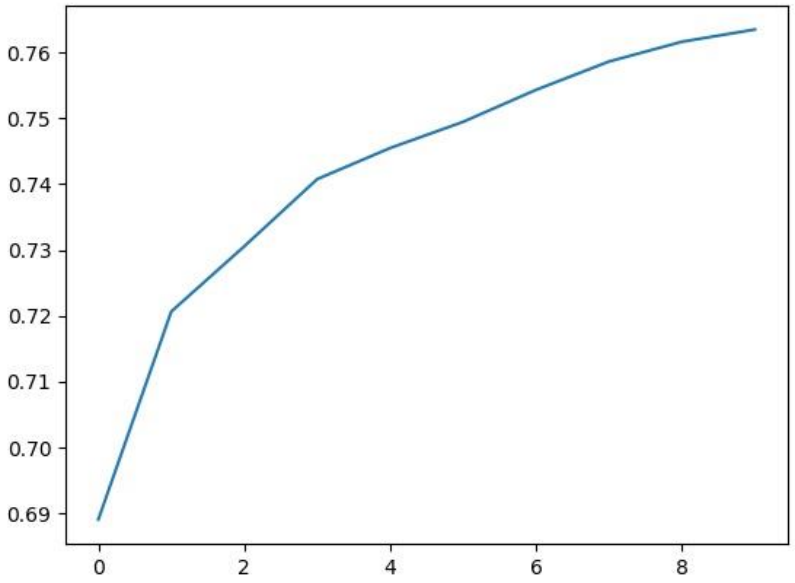
Machine Learning HW6 Report

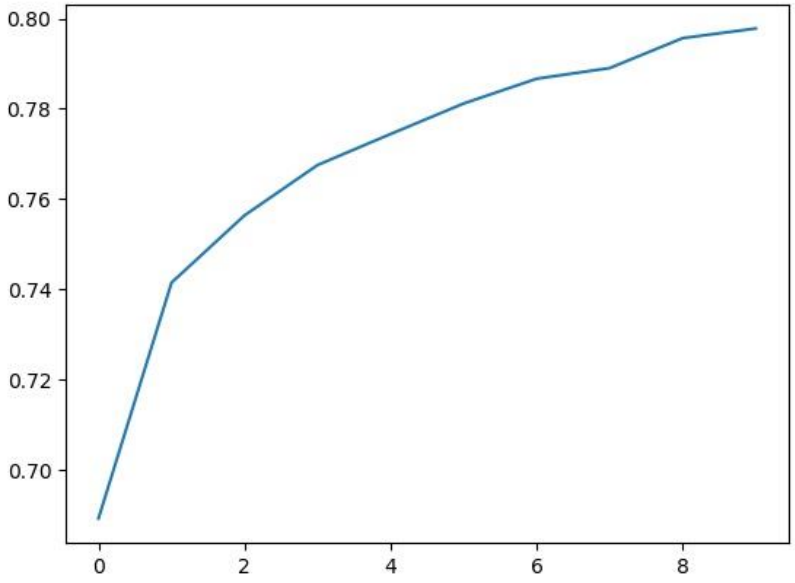
學號：R07922104　　系級：資工碩一　　姓名：林傳祐

1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線*
   - RNN 架構:
     Embedding(word_num,word_dim)　　ps. use word_dim=100
     GRU(word_dim, word_dim,layer_num=2,batch_first=True,dropout=0.7)
     FC(
       Dropout(0.5),
       Linear(word_dim,1),
       Sigmoid(),
)
   - Word embedding 方法:
     GENSIM 套件的 Word2Vector
     word2vec.Word2Vec(sentences, size=100, workers=8, window=3, iter=16, sg=1)
     SG=1 是使用 skip-gram，預設為 CBOW
   - ACC=0.74530 (KAGGLE)



2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。
   - BOW+DNN 架構:

     FC(
       Linear(word_num,16), ps. Total 121756words
       ReLu(),
       Dropout(0.5),
       Linear(16,1),
       Sigmoid(),
)
   - ACC=0.73380 (KAGGLE)



3. (1%) 請敘述你如何 improve performance（preprocess, embedding, 架構等），並解釋為何這些做法可以使模型進步。
   - Preprocess: 我覺得有些表情符號 or 顏文字會影響 Model 的判斷，因此使用 RE 將他們過濾掉
   - Embedding: embedding 也一起 Train 能做更好，因為之後的 Model 在 train 時，能比一開始 Word2Vectord 看得更多，進而使 embedding layer 更好
   - Dropout,Regularization: RNN 非常容易 overfitting，這兩個方式能避免 model 過度 overfit

4. (1%) 請比較不做斷詞 (e.g., 以字為單位) 與有做斷詞，兩種方法實作出來的效果差異，並解釋為何有此差別。
   - (無斷詞)0.71330 vs 0.74530
   - 以中文來講，一個字不一定能表達整個詞的意思，因此不做斷詞，Model 可能無法真正學到詞與詞之間的意義，導致無法做的好
5. (1%) 請比較 RNN 與 BOW 兩種不同 model 對於 "在說別人白痴之前，先想想自己" 與"在說別人之前先想想自己，白痴" 這兩句話的分數（model output），並討論造成差異的原因。
   - RNN: 0,1
   - BOW: 1,1
   - 兩個句子中，”白痴”這個字絕對是判斷的依據。
     在 RNN 中，Model 能同時學到前後文，因此判斷”白痴”時，還有其他依據，因此在第一句就不會被判斷成惡意言論
     相對地在 BOW 中，只是統計各個詞出現次數，那兩個句子都統計到了”白痴”，那 Model 就很可能把兩句都判斷成惡意言論了