學號：R05942017 系級： 電信碩二 姓名：黃梓明

(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: r05942148 鄭立晟)

答：我先用 gensim.word2vec(min_count=1)建立一個字的模型，資料是利用 trainlabel 和 unlabel 的 data，維度為 100，共約有 24 萬個詞→再將 train label data 句子轉成詞向量→input 到 keras 的 model 裡，input_shape=(50,100),結構為 LSTM(256,return_sequences=True)→LSTM(256)→Dropout(0.5)→Dense(512, activation='relu')→Dropout(0.5)→Dense(256, activation='relu')→Dropout(0.5)→Dense(128, activation='relu')→Dropout(0.5)→Dense(64, activation='relu')→Dropout(0.5)→Dense(32, activation='relu')→Dense(1, activation='sigmoid')，

loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'], batch_size=512, epochs=30
callbacks = [
EarlyStopping(monitor='val_acc', patience=30, verbose=1),
ModelCheckpoint(filepath, monitor='val_acc', save_best_only=True, verbose=1),
ReduceLROnPlateau(monitor='val_loss', factor=0.1, patience=10, verbose=0, mode='auto', epsilon=0.0001, cooldown=0, min_lr=0)]
上傳到 kaggle 上的準確率為 81.240%

(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？

答：我一樣先用 gensim.word2vec 建立一個字的模型，但由於記憶體不足的關系，詞不能太多，設 min_count=100，沒有符號，詞的數量約為 7200 個，再將這 7200 用 gensim.dictionary 建一個字典，input_dim=72**，結構為 Dense(512, activation='relu')→Dropout(0.5)→Dense(256, activation='relu')→Dropout(0.5)→Dense(128, activation='relu')→Dropout(0.5)→Dense(64, activation='relu')→Dropout(0.5)→Dense(32, activation='relu')→Dense(1, activation='sigmoid')，

loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'], batch_size=512, epochs=30
callbacks = [
EarlyStopping(monitor='val_acc', patience=30, verbose=1),
ModelCheckpoint(filepath, monitor='val_acc', save_best_only=True, verbose=1),
ReduceLROnPlateau(monitor='val_loss', factor=0.1, patience=10, verbose=0, mode='auto', epsilon=0.0001, cooldown=0, min_lr=0)]
基本上就是第一題的結構再拿掉 LSTM，上傳到 kaggle 上的準確率為 79.525%，但記憶體用量比 RNN 大很多

(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。

答：對於(a) today is a good day, but it is hot,RNN 的 pred 為 0.717989, BOW model 的 pred 為 0.680047

(b) today is hot, but it is a good day,RNN 的 pred 為 0.95763 ,BOW model 的 pred 為 0.680047

對於這兩者的差別，RNN 會參考字的先後順序，故 input 會不一樣，而 BOW 不會參考先後順序，input 一樣，所以 BOW 的 pred 會一樣

(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。

答：”有”標點符號的準確率為 81.240%，”無”的則是 80.612%，我覺得在準確率上差別很小，而詞的 model 在沒有標點符號的情況下可以縮小，若是資料量在比現在的大個十倍百倍，我覺得有無標點符號沒有影響，而沒有標點符號的 model 可以變小

(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。

答：由於電腦記憶體不夠大，我將詞向量的維度從 100 降低到 10，上傳到 kaggle 的準確率從 81.24%降到 77.625%

我把 nolabel 的資料直接丟進模型裡預測，取 label 大於 0.75 當 1，小於 0.25 的當 0，剩下的丟棄不用，約有 90 萬筆可用，再將這個 label 好的資料加入 train data 裡重新 Training，training 時的 val_acc 到 95%，而上傳到 kaggle 上的準確率為 78.151%，有微幅上升 78.151-77.625=0.526%