學號：R06922095 系級： 資工碩一 姓名：陳代穎

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: )

答：模型架構為Layer (Output Shape): Input layer (, 40) -> Embedding layer (, 40, 300) -> Bidirection GRU layer (, 40, 1024) -> GRU layer (, 512) > Dense layer (, 256, relu) -> Output layer (, 1, sigmoid)。其中Input layer長度為長40的sequence，Embedding layer利用gensim.model.word2vec pre-trained出embedding matrix。訓練過程之loss function為binary crossentropy，optimizer為adam，batch size為128，train 10個epochs，利用checkpoint存validation accuracy最高的model。最終在Kaggle上的準確率為public: 0.81709、private: 0.81724。

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: )

答：模型架構為Layer (Output Shape): Input layer (, 300) -> Dense layer (, 256, relu) -> Output layer (, 1, sigmoid)。其中dictionary size為300，因為太大時記憶體會不夠，因此BOW的input為300維。訓練過程之loss function為binary crossentropy，optimizer為adam，batch size為128，train 10個epochs，利用checkpoint存validation accuracy最高的model。最終在Kaggle上的準確率為public: 0.70251、private: 0.70366，明顯低於RNN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: )

答：BOW對於2句的word vector都是相同的，因為BOW不考慮詞的順序，2句的output均為0.59421，符合ground truth。RNN對於”today is a good day, but it is hot”的prediction為0.16337，對於”today is hot, but it is a good day”的prediction為0.97947。因為RNN考慮詞的順序，其中”but”這個轉折語會導致句子的語意著重在後半段，猜測”hot”這個詞的情緒偏負面，而”good”的情緒偏正面。導致在整個句子下，前者偏負面，後者則偏正面。

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: )

答：利用同1. 中的RNN架構比較不同tokenize的方式，我利用keras.preprocessing.text的Tokenizer，default的filters參數會濾掉所有標點符號，將參數設為filters=''，即不濾掉標點符號。最終"不含標點符號"的準確率如同1. 中public: 0.81709、private: 0.81724，而"含標點符號"的準確率為public: 0.82325、private: 0.82116。"含"標點符號的準確率高於"不含"標點符號的準確率，推測可能原因為標點符號中，例如：驚嘆號"!"，有強調語氣的作用，因此不論是正面或負面，都能再把情緒推向該方向，使得分數更為兩極。而問號"?"，則能將分數拉回中立，而得到更精準的判斷。

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: )

答：semi-supervised我選擇最簡單的self-training，首先利用labeled data trained好一個model，並用這個model去predict unlabeled data，我設threshold為0.1，表示prediction低於0.1或高於0.9時我才標記label。當每回合標記完後我會累計label的次數，若回合間predict label有變換，則將其累計次數歸零，代表prediction不穩定。當label的累計次數達3次後我再將其加入training set，並繼續train model。最終準確率為public: 0.82305、private: 0.82314，與supervised learning的準確率(同1. )高出public: 0.00595、private: 0.0059。有較高的準確率，但semi-supervised要注意許多細節，如threshold的設定、unlabeled data加入training set的方式…等，否則model很容易overfitting 在training set上。