**Data Mining HW2 Report**

**Student ID: 110550014 Name: 吳權祐**

1. **How do you select features for your model input, and what preprocessing did you perform to review text?**

I used the following code to preprocess the data. I removed the URL, emoji, html labels and all punctuations in the sequence, then I transformed it into lowercase and lemmatized it. Additionally, I added a classify token in front of the sequence and randomly add the mask.

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

1. **Please describe how you tokenize your data, calculate the distribution of tokenized sequence length of the dataset and explain how you determine the padding size.**

The figure below illustrates the distribution of tokenized sequence lengths. The blue curve represents the distribution for the training data, with an average length 27, while the orange curve depicts the distribution for the validation data, with an average length 31. Most sequences have a length of less than 200, so I set the padding size at 256.

一張含有 文字, 圖表, 螢幕擷取畫面, 繪圖 的圖片

自動產生的描述

1. **Please compare the impact of using different methods to prepare data for different rating categories.**

I used three different methods to prepare the data. The results indicate that the model trained on both the title and text data achieves higher accuracy across all rating categories. Additionally, the model trained separately on title data and text data demonstrates comparable performance. Notably, the title-trained model shows increased accuracy for rating 3, while the text-trained model excels in ratings 4 and 5.

**一張含有 文字, 軟體, 多媒體軟體, 螢幕擷取畫面 的圖片

自動產生的描述一張含有 文字, 軟體, 多媒體軟體, 螢幕擷取畫面 的圖片

自動產生的描述一張含有 文字, 軟體, 多媒體軟體, 螢幕擷取畫面 的圖片

自動產生的描述**Only title Only text Title + text