Data Mining HW2 Report　0816077 陳柏安

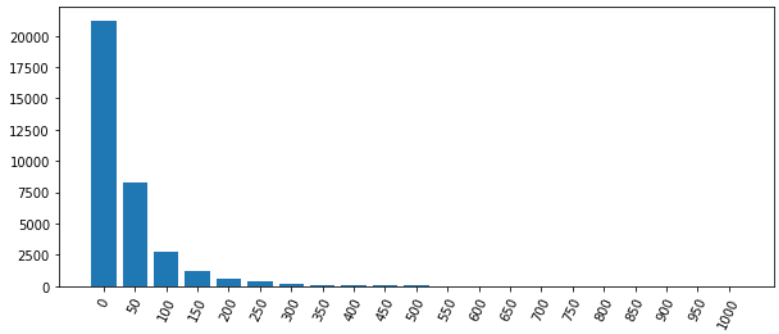
1. How do you select features for your model input, and what preprocessing did you perform to review text?

Feature selection 的部分我簡單的選擇了評論的標題以及內文，並將他們 concatenate 成一整段文字，除此之外並沒有做任何其他處理。

2. 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.

我使用現有的 bert-base-uncased 的 tokenizer 將文本轉化成 tokens，並用 max length = 128 來做 padding 和 truncation。雖然我們的模型最多能使用長度 512 的序列，而且加大 max length、減少 truncation 也能減少訊息遺失，可能增加模型預測的準確率，但經過實驗，發現太長的序列當作 input 會讓訓練時間大幅拉長，一個 epoch 需要超過兩小時才能訓練完成。而且經過計算，在 35000 筆訓練資料中，將近 30000 筆文本在 tokenize 後長度都在 100 以下，所以其實以 128 作為 max length 對整體準確率應該影響不大。

```
0 - 49: 21230
50 - 99: 8334
100 - 149: 2720
150 - 199: 1180
200 - 249: 568
250 - 299: 352
300 - 349: 200
350 - 399: 121
400 - 449: 83
450 - 499: 54
500 - 549: 37
550 - 599: 27
600 - 649: 21
650 - 699: 15
700 - 749: 14
750 - 799: 9
800 - 849: 10
850 - 899: 5
900 - 949: 5
950 - 999: 2
1000 or more: 13
```



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

經過比較，標題和內文 concatenate 後的組合能拿到最高的整體正確率。而比較只是用單一項目的兩者，雖然差別不顯著，但在 f1-score 上，只使用標題能較正確的判斷 1 分的評論，只使用內文能較正確的判斷 2 和 3 分的評論。我推測 1 分的評論可能常使用純粹負面的標題所以容易判斷，而 2 或 3 分的評論可能使用更偏中立委婉的標題，或是只在內文中詳細描述優缺點，讓模型難以只從標題判斷整體評分。

使用標題和內文：

	precision	recall	f1-score
0	0.67	0.73	0.70
1	0.49	0.47	0.48
2	0.54	0.55	0.54
3	0.65	0.57	0.61
4	0.77	0.81	0.79
accuracy			0.63

只使用內文：

	precision	recall	f1-score
0	0.64	0.65	0.64
1	0.44	0.44	0.44
2	0.46	0.48	0.47
3	0.57	0.50	0.53
4	0.72	0.77	0.74
accuracy			0.57

只使用標題：

	precision	recall	f1-score
0	0.61	0.68	0.65
1	0.41	0.36	0.39
2	0.46	0.47	0.46
3	0.57	0.49	0.53
4	0.70	0.79	0.74
accuracy			0.56