Data mining HW2 -

Product review rating prediction

109350008 張詠哲

(For all the experiment result, the metric is **accuracy.** Train on 5 epochs, training: test = 8 : 2)

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

- **Feature selection**

For feature selection, I choose all feature (title, text, verified_purchase) except 'helpful_vote'. Because as the following figure shows that the 'helpful_vote' gets a non Pearson correlation value (0.00) to ground truth ('rating'). Therefore, I don't select this feature. As for 'verified_purchase', it get the -0.11. Though it's not very relate to ground truth, but I think it can make some help to the model.

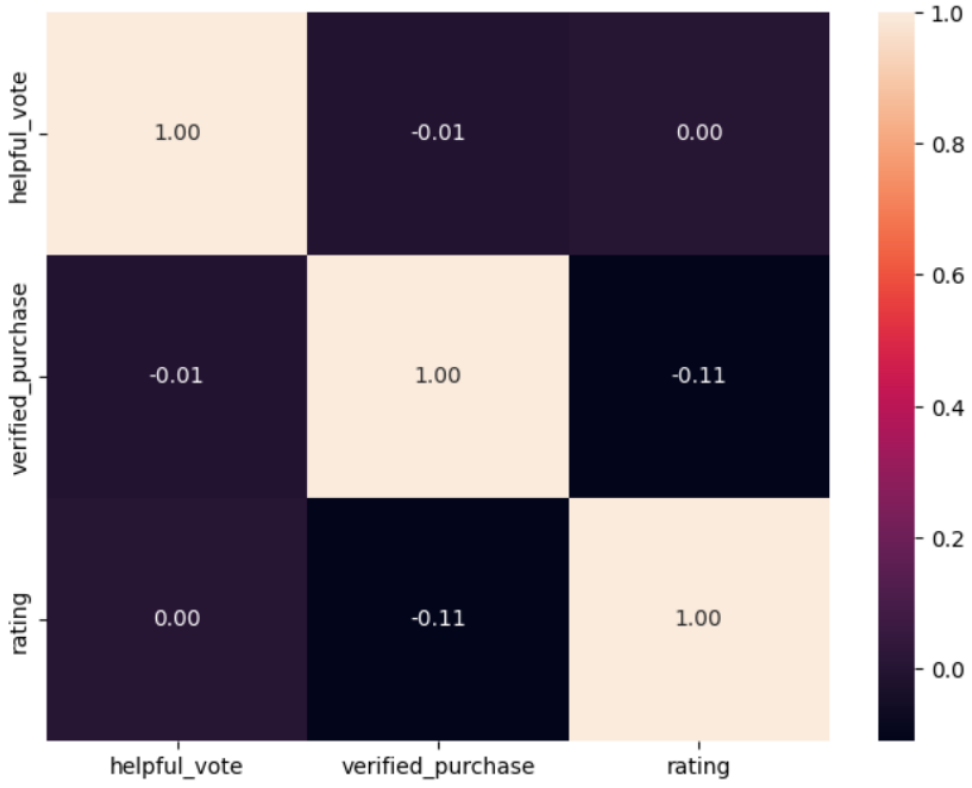

Figure1. Pearson correlation

- **Preprocessing**

For the preprocessing part. I implement 3 methods for review text: remove stop word, remove HTML, lemmatization. Here is a brief introduction about the above

method.

Remove stop word

Remove some prepositions, pronouns, definite articles (e.g., the, who, it, he) and other low-information words.

Ex. “let me do it for you”(input) -> “let”

Remove HTML

I found that the data contain the HTML format, therefore I decide to remove them, such as
</br>

lemmatization

Restore the parts of speech. Ex. happened -> happen

And here is the result of different preprocessing method.

	Accuracy
All	0.594
Stopword only	0.612
HTML	0.606
Lemmatization	0.596
No preprocessing	0.625

Table1

As the table1 shows that, the “no preprocessing” gets the best result. I think the following reasons may cause this result (no processing better than preprocessing).

- Firstly, stop words removal may remove important context information that could be relevant for the model's understanding. While stop words like "the," "and"..., are common and may not carry much meaning individually, they can still contribute to the overall context and semantics of a sentence.
- Second, HTML removal might be beneficial in cases where the HTML tags add noise to the text. However, if the HTML tags contain relevant information, such as formatting cues or specific text attributes, their removal could lead to loss of important data.
- Last, lemmatization aims to reduce words to their base or root form, which can help in reducing vocabulary size and improving generalization. However, in some cases, the lemmatization process might not preserve the exact meaning or context of the words, leading to a loss of specificity in the text.

Therefore, I don’t take any preprocessing for review text. Just use the raw review text.

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.**

I used AutoTokenizer to get the tokenizer from the pre-trained distilled BERT model. This tokenizer can convert text data into a sequence of tokens acceptable to the model. Though distilled BERT is a lighter version of BERT, but it gets a faster computing time while maintaining similar performance. And I think distilled BERT might can less the chance of overtraining.

Here is the distribution of tokenized sequence length in training data.

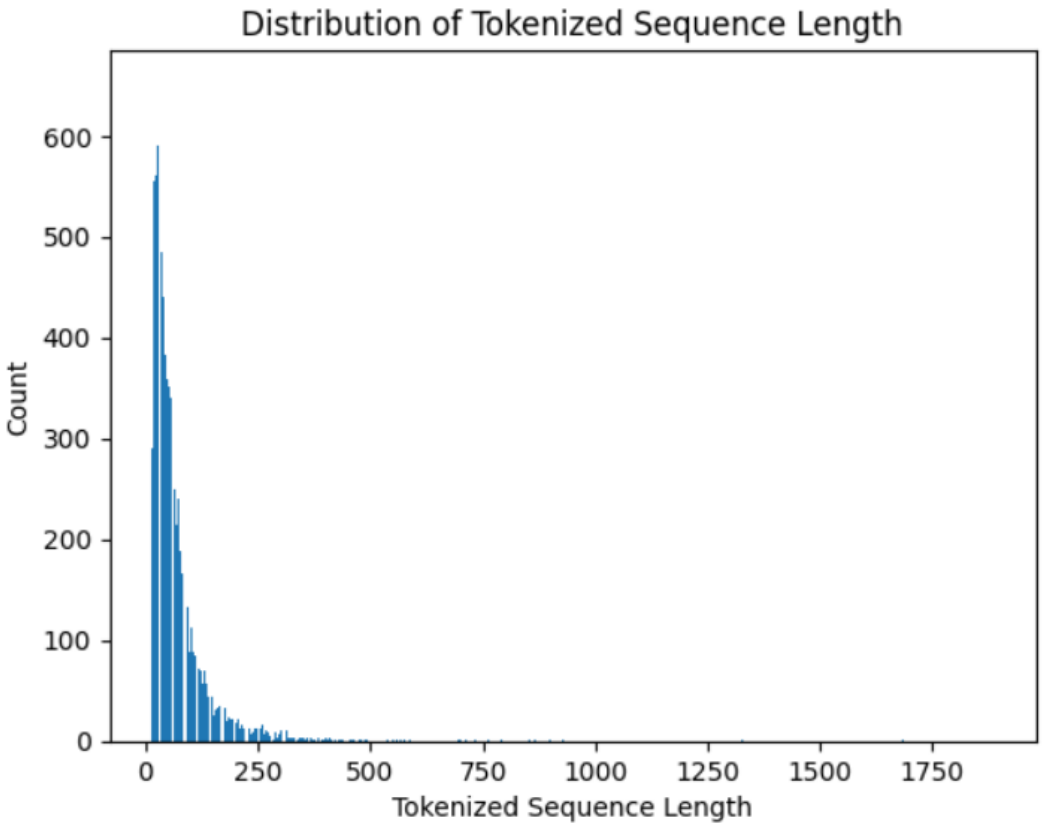

Figure2. distribution of tokenized sequence length in training data

As the above figure shows that most tokenized sequence length is in the 0~250 sequence. Therefore, I pad the data with the 256 padding size. And also do the truncation with same size to increase the computation efficiency and avoid overfitting. This is the tokenize parameter I settled.

```
self.tokenizer(data, return_tensors='pt', padding='max_length', truncation=True, max_length=256)
```

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

In this section, I experiment 4 methods to prepare data by using: title + text + verified_purchase, title + text, title only, text only. And the following figures are the confusion matrix on testing data.

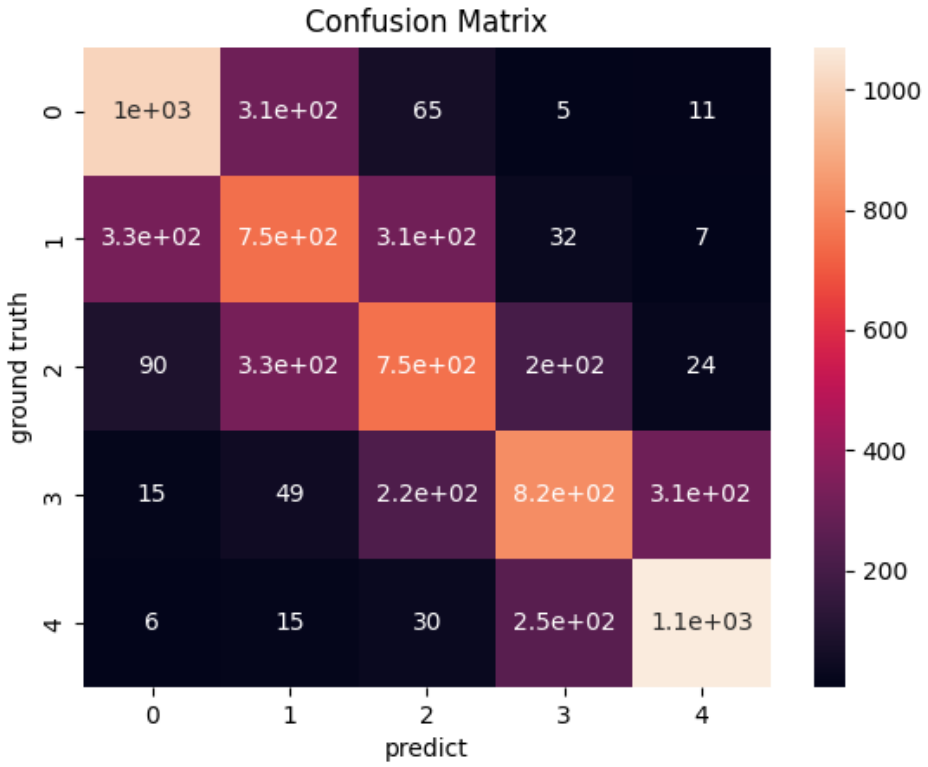

Figure3. title+text+verified_purchase

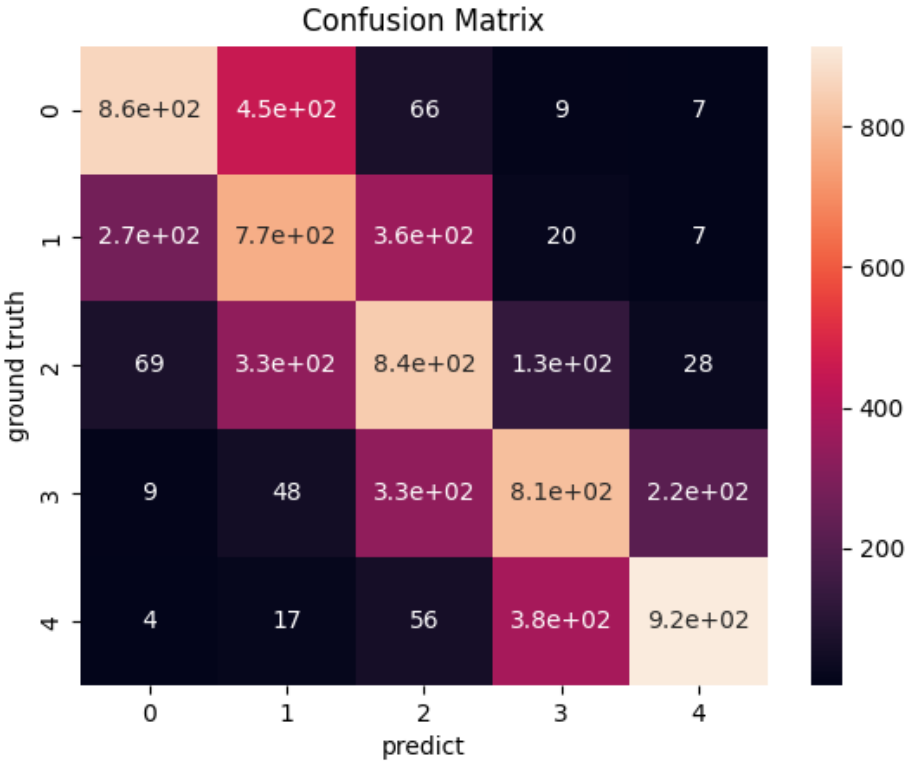

Figure4. title + text

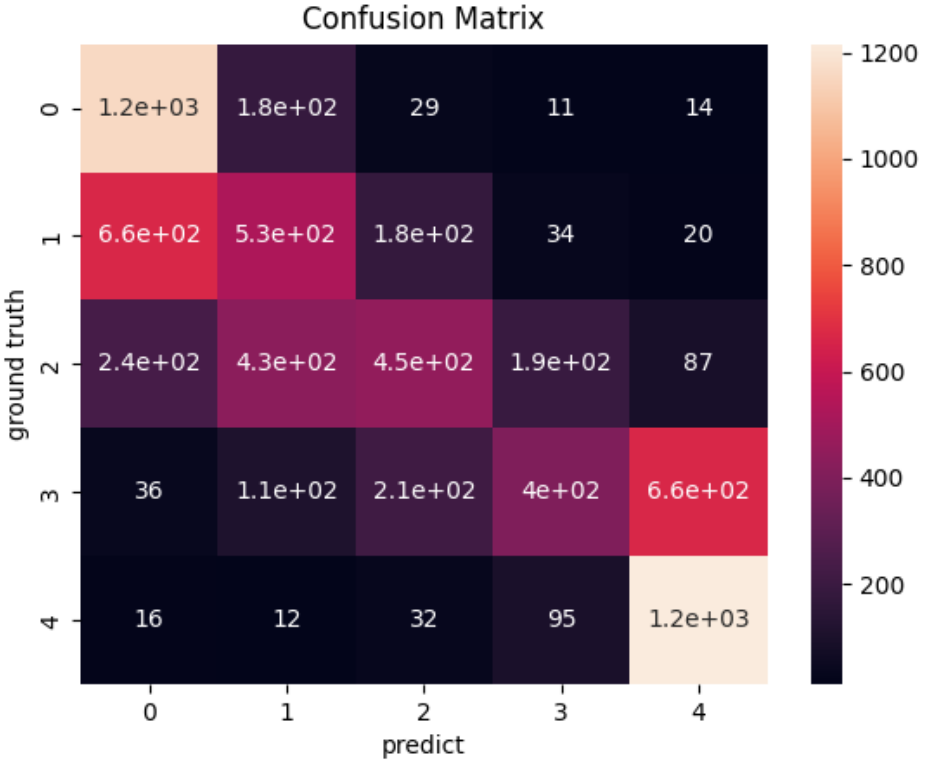

Figure5. title only

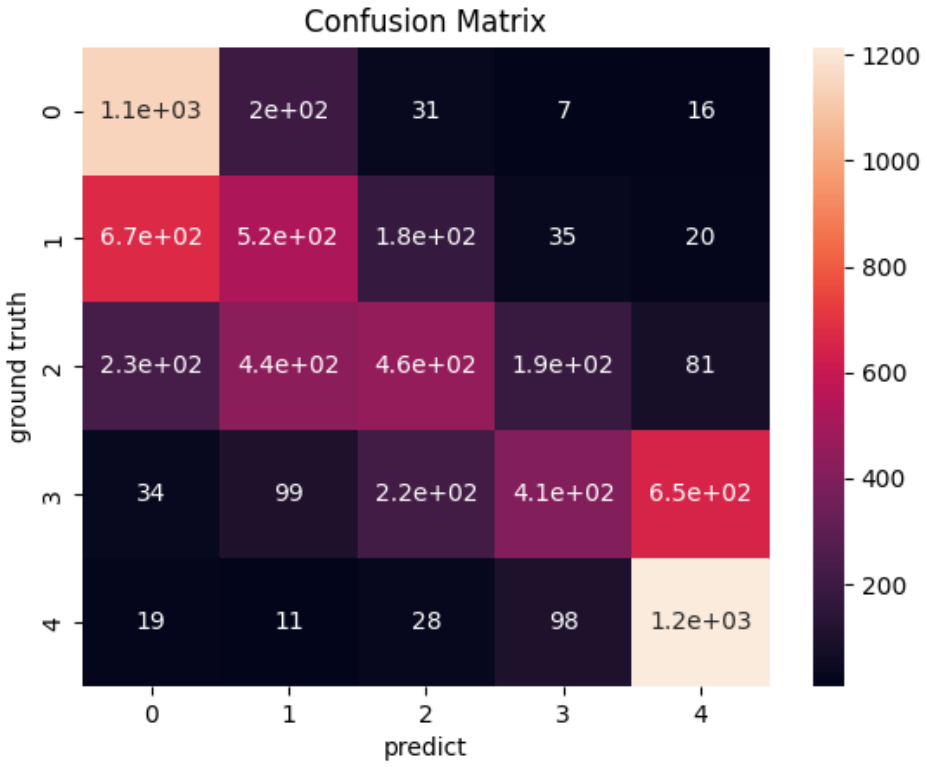

Figure6. text only

And the following table is the best rating accuracy with different prepare methods.

Rating	Method
1	Title only
2	Title + Text
3	Title + Text
4	Title + Text + verified_purchase
5	Title only & Text only

As the above table shows that the best rating accuracy on 1~5 are Title only, Title + Text, Title + Text, Title + Text + verified_purchase, Title only & Text only (same performers). In my opinion. For extreme ratings (1 and 5), title only may capture the essence or sentiment of the review effectively, causing higher accuracy when used as the sole feature. This could be because titles often summarize the main point or emotion of the review. And this result is the most surprise me, I originally think the result by only using title will get the worse result.

In contrast, for intermediate ratings (2, 3, 4), the combination of both title and text provides a more comprehensive view of the review's content, sentiment, and nuances,

leading to better prediction accuracy.

And by adding the "verified_purchase" feature likely improves accuracy for certain rating categories (such as rating 4) where the authenticity or credibility of the reviewer's experience plays a significant role in rating the product.

In conclusion, longer data (combining title and text) may offer more context and details, making them more informative for predicting intermediate ratings. Shorter data (title only) might be more straightforward and sentiment-driven, making them effective for extreme ratings.