DM HW2

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

   I try different ways as the input. At first, I tried to use all the features (title + verified_purchase + text + helpful_vote. It works well with macro f1 0.63, and I tried other ways. According to experiments, title+text+helpful_vote perform the best on the validation set. I split the validation set training set 1:9. I replace the “

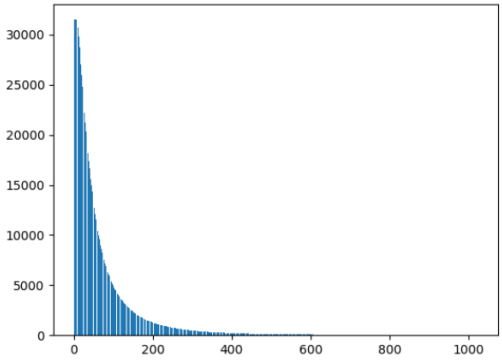
 with single space, “
” with single space, and remove all the things inside ‘[[.....]]’.

```
# replace '[[...]]' as ''
data = re.sub(r'\[\[.*\]\]', '', data)

data = data.replace('<br /><br />', ' ').replace('<br />', ' ')
```

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 use the tokenizer based on the pretrained model I use. I checked the tokenized length for all the data using titles and comment. The graph below is the histogram to visualize the token lengths of all data. The x coordinate is the token length, and the y axis are the number of data points having token bigger or equal to x. Based on the following graph set the maximum length of the tokens to 256, striking a balance between the training time and truncated tokens.



The actual configuration:

```
output = 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

   I recorded macro F1 scores of 4 types per epoch, all feature / only text / only title / text + title. You can see that the one with full features and the one with text and title perform the best.