2024 INLP HW1

110550085 房天越

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

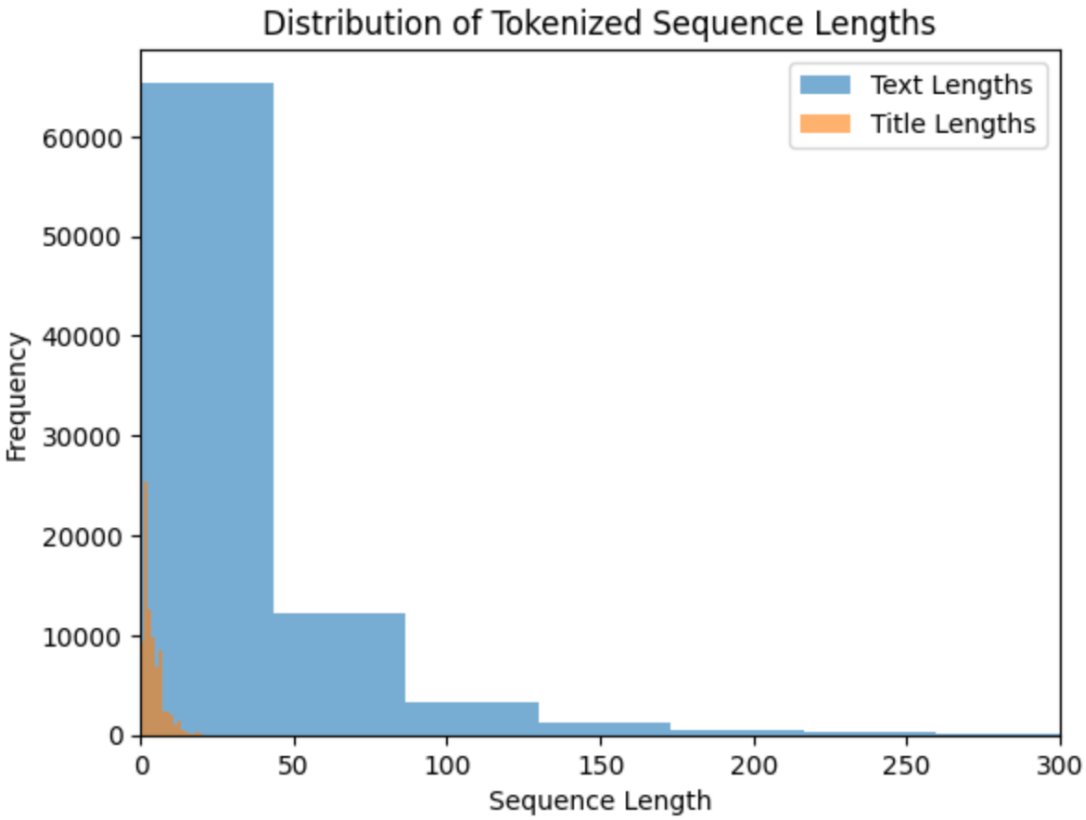
I select the title and the text columns from the dataset, both are textual data, and would be processed into numerical form through tokenization and embedding layers. Also, helpful_vote is considered for data augmentation, if helpful vote is high enough, it would not have the augmentation in order to highlight their influence.

The processing includes lowercasing, tokenization, padding, and data augmentation. For lowercasing, all text is converted to lowercase to maintain the consistency. For tokenization, it’ s performed on both text and title, where each word is converted into a sequence of integers, representing a unique token in the vocabulary. For padding, it is to ensure that all input sequences are of the same length. Finally, for augmentation, it includes synonym replacement, random insertion, and deletion to augment the data, and highlighting the helpful reviews.

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 from Keras to convert the text and titles into sequences of integers, the tokenizer is trained on the training data.

The distribution of tokenized sequence length is here:



There are more than 35000 data because there are some augmentation and create new data.

So the padding size is determined by the largest length, which is about 300 for text, and 20 for title.

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

   I. Augment or not

      Without augmentation, data with fewer helpful votes or underrepresented categories suffer from worse performance. With augmentation, you can increase the amount of training data, have a better generalization, and have better performance.

   II. Use Embedding or train from scratch

      If we train from scratch, the speed would be faster, but also less powerful. With pretrained embeddings like word2vec, we have meaningful word representation because they are trained on large corpora, which leads to improved performance.

   III. Impact on different categories

      Data augmentation specifically targeting the less frequent ratings, combined with class weights, can predict them better. Also, after training, we analyze the classification errors by looking at the confusion matrix or F1 scores, this helps adjusting the augmentation and preprocessing techniques for specific rating categories.