NLP HW1

110550108 施柏江

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

   Ans:
   In my model, I selected the following features for input:
   1. Text features: I combined the "title" and "text" of the product review, then converted them to lowercase. I used CountVectorizer to transform the text into a bag-of-words representation with a maximum of 10,000 features.
   2. Verified purchase: I encoded this categorical feature into numerical values using LabelEncoder.
   3. Helpful votes: This numeric feature was included directly.

   For preprocessing the review text, I applied the following steps:
   1. Combined the title and review body.
   2. Converted all text to lowercase.
   3. Used CountVectorizer to convert the text into numerical features by counting word occurrences, limiting the feature set to the top 10,000 most frequent words.

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.**

   Ans:
   In my code, I did not explicitly tokenize the text data. Instead, I used CountVectorizer from scikit-learn, which treats the text as a bag of words, splitting the text into individual words based on spaces. The vectorizer then

converts the text into a sparse matrix of word counts, with a maximum of 10,000 features representing the most frequent words in the dataset. Since I treat the text as bag-of-words features, there is no concept of sequence length or token padding.

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

Ans:
Bag-of-Words: Simply counts word occurrences, treating words independently of each other.
Pros: Fast and simple, often effective for small datasets.
Cons: Does not capture word context or relationships between words, which can limit its ability to differentiate subtle sentiments in reviews.

Word Embeddings: Captures semantic relationships between words and context.
Pros: Often results in better performance by understanding the sentiment or meaning behind reviews, which is crucial for correctly predicting different ratings.
Cons: More computationally expensive and requires larger datasets.