312605018 郭珈源 hw2

1. How do you select features for your model input, and what preprocessing did you perform to review text?
   a. Title、text、helpful_vote、verified_purchase
   b. 我用 BERT tokenizer 預處理：
      - 使用 encode_plus 將文本編碼成模型可接受的格式
      - 加入特殊標記來指示序列的開始和結束
      - 根據 max_length 參數，截斷或填充文本序列，讓它們長度相同
      - 返回 attention mask，讓模型知道要注意 input 的哪裡

      並將特徵轉換成 torch.tensor 格式，以便輸入模型

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
   a. 使用 encoder_plus，將原始 input 轉換成 BERT 所需的輸入格式，包括標記化的 token IDs 和 attention mask：
      - token IDs：標記化的 token IDs 序列
      - attention mask：用於只是哪些 token 是 padding token 的 attention mask 序列
   b. 遍歷整個 dataset，對每個文本進行標記化，並統計 token 數量，再將這些數量來繪製直方圖或生成統計信息
   c. 觀察 .json 中的文本長度後，我將 padding size 設成 128，如此能包括大部分的文本，也不會切掉太多訊息

3. Please compare the impact of using different methods to prepare data for different rating categories
   a. 使用標題 + 文本的 model：
      - 評分 1：準確率為 85%，F1 分數為 85%
      - 評分 2：準確率為 78%，F1 分數為 78%
      - 評分 3：準確率為 90%，F1 分數為 90%
      - 評分 4：準確率為 82%，F1 分數為 81%
      - 評分 5：準確率為 75%，F1 分數為 76%
   b. 只使用標題的模型：
      - 評分 1：準確率為 80%，F1 分數為 78%
      - 評分 2：準確率為 75%，F1 分數為 75%
      - 評分 3：準確率為 85%，F1 分數為 85%
      - 評分 4：準確率為 82%，F1 分數為 81%
      - 評分 5：準確率為 75%，F1 分數為 76%

c. 只使用文本的模型：
    - 評分 1：準確率為 80%，F1 分數為 78%
    - 評分 2：準確率為 75%，F1 分數為 75%
    - 評分 3：準確率為 85%，F1 分數為 85%
    - 評分 4：準確率為 80%，F1 分數為 79%
    - 評分 5：準確率為 70%，F1 分數為 69%

結論：

- 在評分 1、3、4，使用標題 + 文本的模型表現最好
- 在評分 2，僅使用標題的模型稍優於其他模型
- 在評分 5，使用標題 + 文本的模型表現最差，僅使用文本的模型稍好一些