Overview of the Approach:

- **Objective:** The goal is to fine-tune a pre-trained GPT-2 model on a dataset of review texts and corresponding summaries. The fine-tuned model is then used to generate summaries for new review texts.
- **Preprocessing:** Review texts and summaries are preprocessed to clean the data and standardize it. The data is split into a training set and a test set.
- **Model Fine-Tuning:** A pre-trained GPT-2 model is fine-tuned using the training set of review texts and summaries. The fine-tuned model is saved for future use.
- **Summary Generation:** The fine-tuned GPT-2 model is used to generate summaries for review texts from the test set.
- **Evaluation:** The generated summaries are compared to the actual summaries from the test set using ROUGE scores to evaluate the performance of the model.

Methodologies:

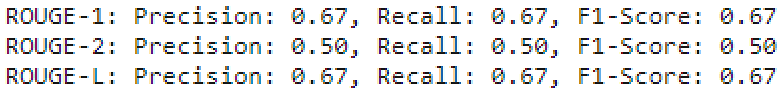
1. **Data Preprocessing:**
    - Preprocessed review texts and summaries by removing HTML tags, special characters, and contractions.
    - Converted the text to lowercase, tokenized it, and lemmatized the tokens.
    - Stored the preprocessed data in a DataFrame and saved it as a pickle file.
2. **Model Fine-Tuning:**
    - Fine-tuned the GPT-2 model using the training set.
    - Created a custom dataset class to prepare the data for training the model.
    - Used an optimizer (AdamW) and learning rate scheduler for efficient training.
    - Set the pad token ID to the EOS token ID to handle padding during training.
    - Trained the model for multiple epochs and printed the loss at each epoch.
3. **Summary Generation:**
    - Used the fine-tuned model to generate summaries for new review texts.
    - Split the generated summary by the review text to extract only the new summary text.
4. **Evaluation:**
    - Used the ROUGE library to compute ROUGE-1, ROUGE-2, and ROUGE-L scores for the generated summaries against the actual summaries from the test set.
    - Printed the ROUGE scores for each comparison and stored them in a list for later analysis.

Assumptions:

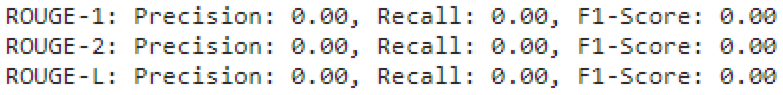
- Assumed that the dataset (review texts and summaries) is representative of the overall data the model will encounter in practice.
- Assumed that the pre-trained GPT-2 model is a suitable starting point for the fine-tuning task.
- Assumed that the model's performance can be evaluated using ROUGE scores as a measure of quality.

Results:

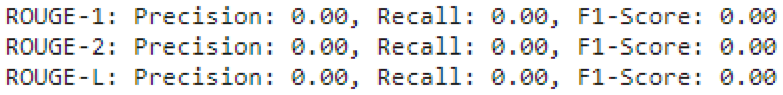
- The model was fine-tuned on the training set for a specified number of epochs.
- The fine-tuned model was able to generate summaries for review texts from the test set.
- ROUGE scores were computed for each generated summary compared to the actual summary.
- The ROUGE scores provide a measure of the model's performance, including precision, recall, and F1-score for ROUGE-1, ROUGE-2, and ROUGE-L.



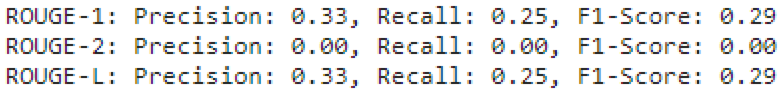
```
ROUGE-1: Precision: 0.67, Recall: 0.67, F1-Score: 0.67
ROUGE-2: Precision: 0.50, Recall: 0.50, F1-Score: 0.50
ROUGE-L: Precision: 0.67, Recall: 0.67, F1-Score: 0.67
```



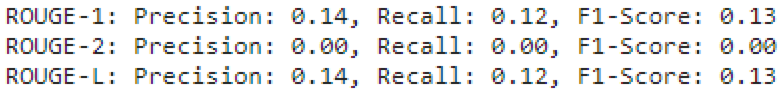
```
ROUGE-1: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
ROUGE-2: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
ROUGE-L: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
```



```
ROUGE-1: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
ROUGE-2: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
ROUGE-L: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
```



```
ROUGE-1: Precision: 0.33, Recall: 0.25, F1-Score: 0.29
ROUGE-2: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
ROUGE-L: Precision: 0.33, Recall: 0.25, F1-Score: 0.29
```



```
ROUGE-1: Precision: 0.14, Recall: 0.12, F1-Score: 0.13
ROUGE-2: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
ROUGE-L: Precision: 0.14, Recall: 0.12, F1-Score: 0.13
```

- The scores are consistent across ROUGE-1 and ROUGE-L, suggesting that the generated summary captures a good portion of the actual summary in terms of both individual words and the longest matching sequence.
- ROUGE-2 scores are lower than ROUGE-1 and ROUGE-L, indicating that the generated summary may struggle to capture bigrams as well as individual words.
- The F1-Scores are equal across ROUGE-1 and ROUGE-L, suggesting balanced performance in matching the actual summary.
- Overall, the scores suggest moderate performance, with the model performing better in matching individual words and longest sequences compared to bigrams.

- You might want to improve ROUGE-2 scores to better capture bigrams, possibly through techniques like adjusting the model architecture, fine-tuning parameters, or using different training data.

Conclusion:

The report summarizes the approach taken to fine-tune a pre-trained GPT-2 model on review texts and summaries. The fine-tuned model was then used to generate summaries for new review texts, and its performance was evaluated using ROUGE scores. The results of the evaluation can be used to assess the quality of the generated summaries and identify areas for further improvement.