1. Assignment 3
2. Report
3. Harsh kashyap
4. 2020434
5.
6. Amazon Fine Food Reviews Dataset: This dataset contains reviews of fine foods from Amazon, including the review text and corresponding summaries.
7. Data Preprocessing: Before training the model, the data needs to be cleaned and preprocessed. This involves tasks like removing special characters, converting text to lowercase, and tokenizing the data to prepare it for training.
8. Model Training:
    - GPT-2 Model: GPT-2 is a state-of-the-art language model developed by OpenAI, capable of generating human-like text.
    - Tokenization: Words or subwords in the text are converted into tokens, which are numerical representations understandable by the model.
    - Custom Dataset Class: This class is implemented to handle the dataset, including tokenization, padding sequences, and preparing the data for training.
    - Fine-Tuning: The GPT-2 model is fine-tuned on the review dataset to generate summaries. Fine-tuning involves updating the pre-trained parameters of the model using the dataset specific to the task at hand.
    - Hyperparameter Tuning: Different hyperparameters such as learning rate, batch size, and number of epochs are experimented with to optimize the model's performance.
9. Evaluation:
    - ROUGE Scores: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate the quality of summaries by comparing them to reference summaries. ROUGE scores measure the overlap between the generated summary and the actual summary in terms of n-grams or sequences of words.
    - ROUGE-1, ROUGE-2, and ROUGE-L: These are different variants of ROUGE scores that focus on unigram, bigram, and longest common subsequences respectively.
    - Precision, Recall, and F1-Score: These are standard evaluation metrics used to measure the quality of generated summaries compared to the actual summaries.

Introduction:

The project aimed to train a GPT-2 model for review summarization using the Amazon Fine Food Reviews dataset. This report outlines the process of data preprocessing, model training, and evaluation of the model's performance using ROUGE scores.
Dataset and Preprocessing:

The Amazon Fine Food Reviews dataset was cleaned and preprocessed by removing special characters, converting text to lowercase, and tokenizing the data to prepare it for training.
Model Training:

- Initialized a GPT-2 tokenizer and model from Hugging Face.
- Split the dataset into training and testing sets with a 75:25 ratio.
- Implemented a custom dataset class to handle the data preparation for training, including tokenization and padding sequences.
- Fine-tuned the GPT-2 model on the review dataset, experimenting with hyperparameters such as learning rate, batch size, and number of epochs.