Documentation for Generating Synthetic Reviews Dataset

The main objective of this study is to implement the GPT-2 model to craft product reviews. The dataset used for estimation was refined and preprocessed, after which the model was tuned to deliver reliable and robust outputs. The resulting product serves as a straightforward instrument for product assessment, although it could be improved with more extensive datasets, clearer models, and enhanced evaluation methods

1. Why was the model/architecture used?

Ans. We chose GPT-2 for this project due to its excellent performance in natural language generation tasks. GPT-2 is a pre-trained transformer-based model recognized for producing coherent and contextually appropriate text. It is especially suitable for generating synthetic reviews, as it is subtle and can grasp patterns present in product reviews, yielding realistic and varied outputs. Furthermore, GPT-2 is adaptable and straightforward to fine tune on specific datasets, making it perfect for this assignment.

2. What were the different factors considered for generating this dataset? (Length, topic diversity etc.)

Ans. Several factors were considered when creating the synthetic review dataset:

- **Length:** The length of each review was limited to 128 characters to maintain conciseness when providing feedback
- **Significance:** This model was developed using a wide range of product reviews, including a diversity of products and feedback types, so that the synthetic data set captures a wide range of themes and perspectives.
- **Accurate methods:** The methods used to generate synthetic data have been refined to ensure high quality and faithful production.
- **Randomization:** In order to randomize and ensure diversity in the studies, top-k and top-p sampling techniques were used during publication.

3. How to measure the performance of a synthetic data set?

Ans. To evaluate the quality and effectiveness of the generated ink dataset, we used quantitative and qualitative methods:

- **Quality evaluation:** Human evaluators work to read a sample of synthetic evaluations were made and based on them. Master in context, communication, creativity and environment. The feedback was collected in order to understand the quality of the reviews made.

- **Quantity evaluation:**

1. **BLEU Score:** We calculate the BLEU score to compare generated reviews with actual reviews and provide a measure of similarity.
2. **ROUGE score:** We use ROUGE criteria to assess the recall of grammars between generated and real reviews, and to evaluate the comprehensibility of the generated text.
3. **Diversity and uniqueness:**We analyze the diversity of comments generated by looking at the percentage of unique comments. We ensure that the generated objects did not simply repeat sentences from the training dataset which increased their quality and usage.

4. How to ensure that the synthetic dataset created is influenced by the source dataset, but not an exact copy?

Ans. To ensure that the synthetic dataset is inspired by the source dataset and not a direct copy:

- **Data cleaning:** Cleaned and pre-processed with inclusion checks to remove duplicate and unnecessary data that's right.
- **Sampling techniques:** top-k and top-p sampling are used during construction to randomize and prevent the model from simply repeating the training data.
- **Diverse Applications:** Uses a variety of messages taken from validated reviews to inspire creativity in the production process and avoid redundancy.

5. What are the main obstacles to solving the problem?

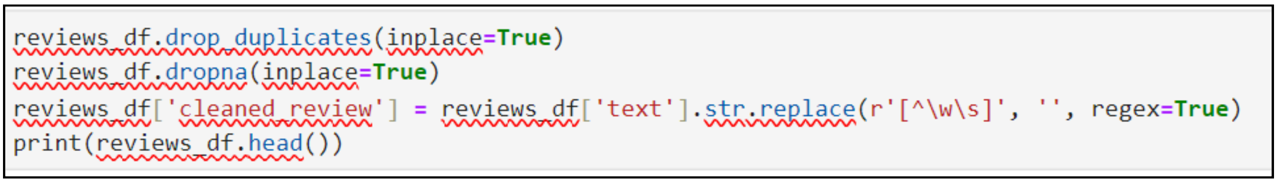
Ans. Several problems arose during the project:

- **Data Quality:** Cleaning the dataset to ensure it was diverse and high-quality took a lot of effort.
- **Storage Issues:** Due to limited storage, we had to reduce the dataset and limit the number of training steps. Storing training and validation losses also requires a lot of space .
- **Realism vs. Diversity:** It was tricky to generate reviews that were both realistic and different from each other. We had to fine tune the model carefully to avoid repetition, spaces and special characters if any.
- **Training Losses:** Tracking training and validation losses takes up a lot of memory and storage making the process more demanding
- **Paid LLMs:** Many advanced language models are paid, which can limit access to more accurate models. Using the free version of GPT-2 might affect the accuracy of the reviews.

Methodology

Step 1 - Data Collection & Cleaning:

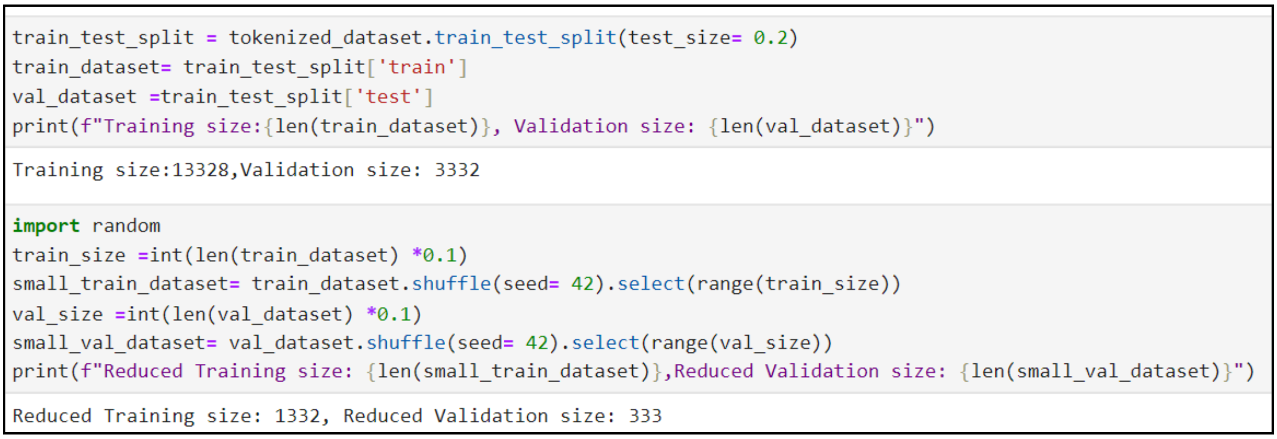
- The product reviews and asins dataset was uploaded, we removed duplicates, dealt with NaN values and normalized the review content by removing special characters with the help of Regular Expressions to ensure quality input.



```
reviews_df.drop_duplicates(inplace=True)
reviews_df.dropna(inplace=True)
reviews_df['cleaned_review'] = reviews_df['text'].str.replace(r'[^\w\s]', '', regex=True)
print(reviews_df.head())
```

Step 2 - Dataset Size Management:

- In order to save space, we opted for only 10% of the dataset to be used for the training and validation processes without necessary memory related problems during the model training.



```
train_test_split = tokenized_dataset.train_test_split(test_size= 0.2)
train_dataset= train_test_split['train']
val_dataset =train_test_split['test']
print(f"Training size:{len(train_dataset)}, Validation size: {len(val_dataset)}")
```

```
Training size:13328,Validation size: 3332
```

```
import random
train_size =int(len(train_dataset) *0.1)
small_train_dataset= train_dataset.shuffle(seed= 42).select(range(train_size))
val_size =int(len(val_dataset) *0.1)
small_val_dataset= val_dataset.shuffle(seed= 42).select(range(val_size))
print(f"Reduced Training size: {len(small_train_dataset)},Reduced Validation size: {len(small_val_dataset)}")
```

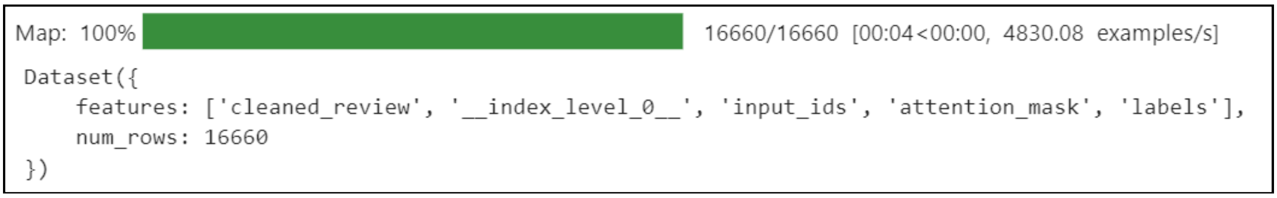
```
Reduced Training size: 1332, Reduced Validation size: 333
```

Step 3 - Model Selection:

- GPT-2 model was made use of since it has great performance on text generation, its comprehension of long range dependencies among languages makes it possible to produce good quality product reviews.

Step 4 - Tokenization:

- The reviews were tokenized using a Hugging Face GPT-2 tokenizer with padding to the end of sequence (eos) token and attention masks for efficient training.

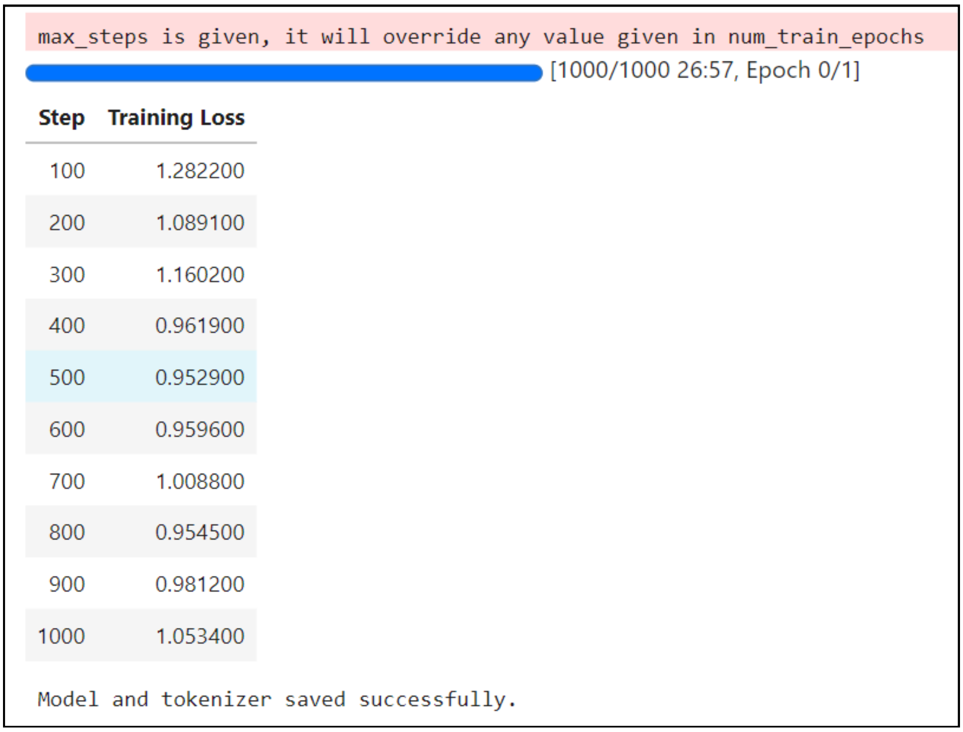


```
Map: 100%  16660/16660 [00:04<00:00, 4830.08 examples/s]

Dataset({
    features: ['cleaned_review', '__index_level_0__', 'input_ids', 'attention_mask', 'labels'],
    num_rows: 16660
})
```

Step 5 - Model Training:

- The training was restricted to 1 epoch and 1000 steps with a low batch size in order to avoid storage issues. The model was tailored to be competent to learn the composition of the product reviews.



```
max_steps is given, it will override any value given in num_train_epochs
```

[1000/1000 26:57, Epoch 0/1]

Step	Training Loss
100	1.282200
200	1.089100
300	1.160200
400	0.961900
500	0.952900
600	0.959600
700	1.008800
800	0.954500
900	0.981200
1000	1.053400

```
Model and tokenizer saved successfully.
```

Step 6 - Synthetic Review Generation:

- Later on and after training, synthetic product reviews were created using tops-k and nucleus top p sampling techniques for diversity. Example of CSV output:

original_prompt	synthetic_reviews	ratings
I have felt better and had more energy taking this	I have felt better and had more energy taking this supplement I could not stand its awful smell I have started taking it and am taking it daily I have lost a couple of pounds My kids are more healthy than me and my younger sister is just getting more and better My toddler has gotten 2 and a half pounds over the past year and my husband is growing into the size he should have been but he has gone back to eating his old way	1
Everything was good i loved it only problem was I	Everything was good i loved it only problem was I wont be ordering again	3
The product is exactly as described and dispenses	The product is exactly as described and dispenses It doesnt have any problems in fact it does seem to be a great quality product but the taste is different from the other brand	5
Great product Tastes fine Much easier than tablets	Great product Tastes fine Much easier than tablets or pills	4
Takes longer than 10 seconds for the temp to show	Takes longer than 10 seconds for the temp to show This is just because it is so small	2

Reasons for using GPT-2 transformer?

- In my consideration, this model was appropriate because of its ability to generate a simple but relevant text in addition to its capturing of complex relationship among languages and this LLM model was a free platform to work in, with an average-to-good accuracy in the prediction model for obtaining the results.

How extensive was the research?

- The research primarily focused on optimizing GPT-2 for generating synthetic reviews, with adjustments to training steps, batch size, and dataset size due to storage limits. We experimented with different sampling techniques, such as top-k and nucleus sampling, to ensure diverse and realistic outputs.

Different factors considered while generating dataset

- While generating the dataset, we considered factors like review length to ensure concise and informative text, topic diversity to capture a wide range of product categories, and randomness. Data quality was also a key to ensure that input reviews were clean and relevant.

Future Improvements

- **Large-Scale Dataset:** The model could benefit from working with unhealthy reviews trained on a larger and more heterogenic dataset.
- **Model Fine-tuning:** Better constraints in the future research would possibly help in obtaining more satisfactory results with the application of different models such as GPT-3, T5, or even better fine-tuning of the GPT-2 on specific product types.
- **Enhanced Evaluation:** The analysis of the evaluation in the review synthesizing systems could be more structured by applying such metrics as sentiment including qualitative assessment of the generated reviews.
- **Real-time Feedback:** Real-time model training whereby there would be human centric intervention over the models and possible feedback on the reviews given could contribute towards a greater degree of accuracy in the diversity of review outputs over time
- **Deployment:** It would be advantageous to deploy the model as a web API or web application whereby reviews generation would be done within real-time context.