Generation of Synthetic Amazon Reviews Using GPT-2

Assignment: Synthetic Data Generation for AI Systems

Submitted By: Vipul Yadav

Email Id: vipul.dtu.2k16@gmail.com

Institution: Delhi Technological University (formerly DCE)

Date: 6th October , 2024

Submitting to: SHL (as assignment)

Abstract

This report explores the generation of synthetic Amazon product reviews using the GPT-2 language model, focusing on supplements and vitamins. The project aimed to create a diverse and realistic set of synthetic reviews to address data scarcity in machine learning applications. By fine-tuning GPT-2 on a dataset of real Amazon reviews and implementing strategies to maintain diversity and relevance, the project successfully generated a synthetic dataset that mimics human-written content while allowing for controlled variations in factors such as review length, sentiment, and product-specific details.

Table of Contents

1. Introduction

In the rapidly evolving landscape of e-commerce and artificial intelligence, the availability of diverse and extensive datasets is crucial for developing robust AI systems. However, obtaining such datasets often poses challenges due to privacy concerns, data scarcity, or high collection costs. This project addresses these challenges by focusing on the generation of synthetic Amazon product reviews for supplements and vitamins using the GPT-2 (Generative Pre-trained Transformer 2) language model.

1.1 Project Objectives

The primary objectives of this project are:

1. To fine-tune GPT-2 on a dataset of real Amazon reviews for supplements and vitamins.
2. To generate diverse and realistic synthetic product reviews.
3. To evaluate the quality and usefulness of the generated reviews using both qualitative and quantitative metrics.
4. To explore the challenges and limitations of using large language models for synthetic data generation.

1.2 Approach Overview

Our approach involves:

1. Preprocessing Amazon reviews data for supplements and vitamins.
2. Fine-tuning GPT-2 on this dataset.
3. Implementing strategies to ensure diversity and relevance in generated reviews.
4. Evaluating the synthetic reviews using metrics like BLEU, ROUGE, and custom diversity scores.

2. Methodology

The project followed these key steps:

1. Data Preparation:
   - Loaded and preprocessed Amazon review data for supplements and vitamins.
   - Cleaned the data by removing duplicates and handling missing values.
2. Model Selection and Setup:
   - Chose GPT-2 as the base model for fine-tuning.
   - Set up the GPT-2 tokenizer with appropriate padding and truncation.
3. Training:
   - Fine-tuned GPT-2 on the preprocessed Amazon reviews dataset.
   - Used a subset of the data (10%) to manage computational resources.
4. Synthetic Review Generation:

    - Developed a function to generate reviews based on prompts from the training data.
    - Implemented strategies to maintain diversity and relevance in generated reviews.
5. Evaluation:
    - Assessed the quality of generated reviews based on fluency, coherence, and relevance.

3. Model Architecture and Justification

While we primarily focused on GPT-2, we considered other architectures:

1. BERT (Bidirectional Encoder Representations from Transformers):
    - Pros: Strong in understanding context.
    - Cons: Not designed for text generation, making it less suitable for our task.
2. T5 (Text-to-Text Transfer Transformer):
    - Pros: Versatile for various NLP tasks.
    - Cons: More complex to set up and potentially overkill for our specific use case.
3. GPT-2 (our chosen model):
    - Pros: Excellent at text generation, easy to fine-tune, and well-suited for review-like text.
    - Cons: Potential for generating less grounded or factual content compared to retrieval-augmented models.

Key reasons for this choice include:

1. Pre-training: GPT-2 is pre-trained on a diverse corpus of internet text, providing a strong foundation for understanding and generating human-like text.
2. Fine-tuning capability: The model can be easily fine-tuned on specific datasets, allowing it to adapt to the particular style and content of Amazon reviews.
3. Context understanding: GPT-2's transformer architecture enables it to capture long-range dependencies in text, crucial for generating coherent and contextually appropriate reviews.
4. Scalability: The model's architecture allows for generating reviews of varying lengths, which is essential for creating a diverse set of synthetic reviews.
5. State-of-the-art performance: GPT-2 has demonstrated excellent results in various text generation tasks, making it a reliable choice for producing high-quality synthetic reviews.

4. Factors Considered in Dataset Generation

Several factors were considered to ensure the quality and diversity of the generated synthetic reviews:

1. Review Length:
   - Set a maximum length of 128 tokens to ensure concise yet informative reviews.
   - Varied lengths to mimic the natural distribution found in real reviews.
2. Topic Diversity:
   - Used a wide range of product names and descriptions as prompts to generate reviews across different supplement and vitamin types.
3. Sentiment Variation:
   - Incorporated different sentiment levels (positive, negative, neutral) by using varied prompts and adjusting generation parameters.
4. Product-Specific Information:
   - Included product names and key features in generation prompts to ensure relevance.
5. Language Style:
   - Fine-tuned on actual Amazon reviews to capture the typical language style used by customers.
6. Rating Consistency:
   - Generated ratings (1-5 stars) along with the review text to maintain consistency between sentiment and numerical rating.
   - vitamins to avoid spreading misinformation.

5. Measuring Efficacy of the Synthetic Dataset

To evaluate the quality and effectiveness of our synthetic Amazon reviews dataset, we employed a combination of quantitative metrics and qualitative assessments. This multi-faceted approach allowed us to gauge various aspects of the generated reviews, including their similarity to real reviews, linguistic quality, and potential usefulness for downstream tasks.

5.1 Quantitative Metrics

5.1.1 BLEU Score

- **Method**: We calculated the BLEU (Bilingual Evaluation Understudy) score to measure the similarity between our synthetic reviews and the original prompts used to generate them.
- **Result**: Our synthetic reviews achieved a BLEU score of 0.4494.
- **Interpretation**: This moderate score indicates that our synthetic reviews capture a significant portion of the information or style present in the original prompts, while also

introducing novel content. It suggests a good balance between staying true to the source and generating new, creative text.

5.1.2 ROUGE Scores

We computed various ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores to assess the overlap between synthetic reviews and original prompts:

a) ROUGE-1: 0.6236

- **Interpretation**: About 62.36% of individual words in the synthetic reviews are also found in the original prompts, indicating good content overlap at the unigram level.

b) ROUGE-2: 0.6033

- **Interpretation**: This score suggests substantial bigram overlap, demonstrating that our synthetic reviews maintain much of the phrasing and context of the original prompts.

c) ROUGE-L: 0.6236

- **Interpretation**: This score reflects that our synthetic reviews preserve much of the original structure, indicating coherent sentence construction.

5.1.3 Diversity Score

- **Method**: We calculated a diversity score to measure the uniqueness of vocabulary across the synthetic reviews.
- **Result**: Our synthetic reviews achieved a diversity score of 0.4255.
- **Interpretation**: This moderate score indicates that while our synthetic reviews contain a good variety of words, there's still some repetition. It suggests room for improvement in increasing the diversity of language used.

6. Ensuring Originality of Synthetic Data

To ensure that the synthetic dataset was inspired by but not an exact replica of the source data, several strategies were implemented:

1. Prompt Engineering:
   - Used diverse and partial prompts from the training data to inspire generation without copying.
2. Temperature and Top-k Sampling:
   - Adjusted the temperature and used top-k sampling during generation to introduce randomness and creativity in the output.
3. Post-processing Filters:

    - Implemented filters to remove any generated reviews that were too similar to those in the training set.

4. Data Augmentation:
    - Applied techniques like synonym replacement and sentence reordering to further diversify the generated content.
5. Combination of Product Features:
    - Generated reviews by combining features from different products to create unique descriptions.
6. Varied Input Lengths:
    - Used input prompts of different lengths to encourage diverse outputs.

7. Challenges and Solutions

Several challenges were encountered during the project:

1. Limited Computational Resources:
    - Challenge: Training on the full dataset was computationally intensive.
    - Solution: Used a subset (10%) of the data for training and implemented efficient batching.
2. Maintaining Review Quality:
    - Challenge: Some generated reviews lacked coherence or product relevance.
    - Solution: Implemented more stringent filtering and used product-specific prompts to improve relevance.
3. Avoiding Repetition:
    - Challenge: The model sometimes generated repetitive phrases or reviews.
    - Solution: Adjusted sampling parameters and implemented diversity penalties in the generation process.
4. Balancing Creativity and Accuracy:
    - Challenge: Highly creative reviews sometimes lacked factual accuracy about products.
    - Solution: Fine-tuned the balance between temperature settings and top-k sampling to maintain creativity while preserving product-specific information.
5. Ethical Considerations:
    - Challenge: Ensuring generated reviews did not contain biased or inappropriate content.
    - Solution: Implemented content filters and conducted manual reviews of samples to maintain ethical standards.