Synthetic Data Generation for AI Systems

Daksh Dudeja

Software Developer, SHL

5st October 2024

INTRODUCTION

This research assignment aims to create a methodology for generating synthetic datasets of Amazon product reviews. Given the extensive amount of user-generated content on e-commerce platforms like Amazon, the ability to simulate realistic review data offers significant potential for applications such as sentiment analysis, recommendation systems, and consumer behavior studies. The synthesis of artificial review data is designed to achieve the following objectives:

a. **Faithfully capture** the linguistic patterns, content structure, and sentiment variability typical of authentic Amazon consumer feedback.
b. **Protect individual privacy** by eliminating the need for actual customer data in analytical processes.
c. **Establish a scalable framework** for generating a wide range of simulated reviews, suitable for various research and development applications.

METHODOLOGY

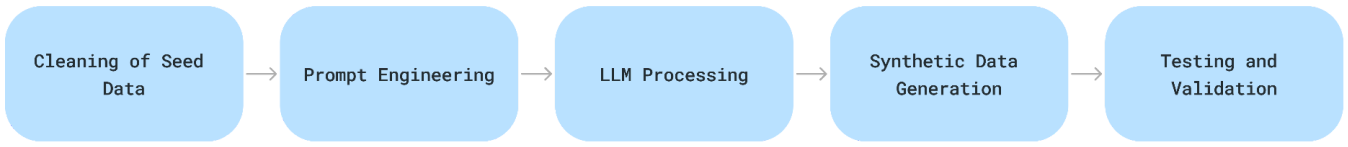

Figure 1: Approach to Synthetic Dataset Generation

Step 1: Seed Data Preparation

I incorporated two datasets into my analysis. The first is a subset of the Amazon reviews dataset, which contains reviews from buyers specifically for the Supplements/Vitamins product category. This dataset includes fields such as **`rating`**, **`title`**, **`text`**, **`parent_asin`**, **`user_id`**, **`timestamp`**, **`asin`**, **`helpful_vote`**, **`verified_purchase`**, **`date`**, and **`time`**. The second dataset comprises **`product_name`**, **`parent_asin`**, and the **`category`** type of each product. Below are the steps taken for data cleaning.

a. **Merged Datasets**: Combined the two datasets to associate product names with their parent ASINs, leveraging Amazon's ASIN assignment during product creation.
b. **Removed Redundant Column**: Eliminated the **`asin`** column as it duplicated the information in the **`parent_asin`** column.
c. **Dropped Verified Purchase Column**: Removed the **`verified_purchase`** column since all entries indicated positive verification.

	A	B	C	D	E	F	G	H	I	J
	rating	title	text	parent_asin_x	user_id	timestamp	helpful_vote	product_name	categories	cat1
	4	B Complex in ge	I bought this alor	B00012ND5G	AGDVFFLJWAC	2009-12-11 0:37	1	BComplex 50 – :	['Health & Household'	Heart Health Event']
	5	Five Stars	great product	B00013Z0ZQ	AG3BSKXHDGF	2015-01-04 3:11	0	Twinlab Ultra GL	['Health & Household'	Heart Health Event']
	5	Five Stars	Came as expect	B00013Z0ZQ	AHG2WKFD4LX	2015-09-27 19:1	0	Twinlab Ultra GL	['Health & Household'	Heart Health Event']
	5	Vitamin Shoppe	Excellent Produc	B00013Z1KA	AEOF7RT3AC4	2019-02-09 19:3	0	NonOily Dry A	['Health & Household'	Heart Health Event']
	5	Un producto que	Es muy buena vi	B00013Z1KA	AGW2WETWQF	2022-07-25 14:1	0	NonOily Dry A	['Health & Household'	Heart Health Event']
	5	Natural Sources	I have been takir	B00014D0IY	AHAHKQ4GXW	2015-12-18 14:0	0	Vitamin A D	['Health & Household'	Heart Health Event']
	5	Liver Support su	I have had good	B00014DY62	AHZSW5KOLYC	2014-03-23 17:5	0	Country Life Live	['Health & Household'	Heart Health Event']
	4	I feel this is work	I feel this is work	B00014DY62	AFNIV34LQAGE	2016-05-13 6:14	0	Country Life Live	['Health & Household'	Heart Health Event']
	5	Five Stars	👍	B00014DY62	AGBGE7ELMW	2016-06-29 0:05	0	Country Life Live	['Health & Household'	Heart Health Event']
	5	Five Stars	Very good value	B00014DY62	AHM5DKU6HRI	2016-12-08 20:1	0	Country Life Live	['Health & Household'	Heart Health Event']
	5	Good product	As per expectati	B00014ECTA	AGIPFR5ZH6SA	2020-10-19 19:3	0	COUNTRY LIFE	['Health & Household'	Heart Health Event']
	5	I love this produc	I love this extren	B00014FSO8	AGRBUBT7LHP	2014-09-25 9:46	1	Food Carotene	['Health & Household'	Heart Health Event']
	5	Five Stars	Get all you need	B00014GBEY	AHYFSYXZ3BZI	2016-07-14 11:1	0	Ultra I Sustained	['Health & Household'	Heart Health Event']
	4	Working Wll	Am taking at pre	B00014I4KS	AFREVDPZUNE	2013-11-03 13:5	0	B2	['Health & Household'	Heart Health Event']
	5	Five Stars	Tiny pills that are	B00014UG7W	AHTUYY6KIK7H	2017-08-31 0:23	0	Food Carotene	['Health & Household'	Heart Health Event']
	4	Different Packag	It looks like the p	B000169ELO	AET25RCAQHB	2012-03-10 1:48	4	Nutricology L-Me	['Health & Household'	Vitamins & Supplements']
	1	Aweful smell	These smelled s	B000169ELO	AGYGMBYTNX	2013-05-25 7:13	4	Nutricology L-Me	['Health & Household'	Vitamins & Supplements']
	2	Strong smell & le	I use L-methionii	B000169ELO	AGZNVWN34LU	2013-06-14 18:0	5	Nutricology L-Me	['Health & Household'	Vitamins & Supplements']
	5	Solves a LOT of	If you know what	B000169ELO	AEFQNYK2UW	2013-07-13 20:5	8	Nutricology L-Me	['Health & Household'	Vitamins & Supplements']
	2	Stopped after tal	It's not for me. I	B000169ELO	AGYCDIDUI7CN	2014-07-19 16:3	4	Nutricology L-Me	['Health & Household'	Vitamins & Supplements']
	5	Love this Produc	I love this produc	B000169ELO	AGIBWKUH4SN	2017-05-13 16:0	2	Nutricology L-Me	['Health & Household'	Vitamins & Supplements']
	5	picture is not cor	This is a wonder	B00016AU6W	AEKVJPBQ7ZS	2012-09-30 21:1	1	Zand Allergy Sei	['Health & Household'	Vitamins & Supplements']
	5	Excellent produc	Much more effec	B00016AU6W	AGKCLCBZ2KP	2013-03-09 17:1	0	Zand Allergy Sei	['Health & Household'	Vitamins & Supplements']
	5	Five Stars	Just what I need	B00020IHJO	AECJCSG2WLN	2015-06-03 17:3	0	Nature's Life Gre	['Health & Household'	Vitamins & Supplements']
	5	I take them ever	I've been buying	B00028MWT2	AEC6VRIQHKD	2014-05-09 5:06	0	NutriGenic	['Health & Household'	Heart Health Event']
	5	Five Stars	Purchased for m	B00028NB3I	AGX3B56CGCV	2017-04-08 10:1	0	Greens Plus Ene	['Health & Household'	Energy & Endurance']
		2 6 capsules a da	Didn't know that	B00028NEW0	AELOXX25ISEF	2018-06-14 16:4	0	Douglas Labs	['Health & Household'	Heart Health Event']

Table1: Well-Organized and Preprocessed Seed Data for Effective Utilization

Step 2: Prompt Engineering and LLM Processing

LangChain is chosen as the framework to handle LLM interactions because it provides a structured way to manage prompt templates, responses, and other components needed to dynamically generate the dataset. It allows for easy integration with LLMs and scales well for multiple steps.

GPT-3.5-turbo serves as the core LLM for generating text due to its high capacity for understanding context and producing human-like text. It is suitable for handling complex review generation by incorporating various factors such as rating, product title, and product category.

Phase A: Random sampling is done for given fields.

a. `p_asin`: Taken directly from the `parent_asin_x` column in the merged dataset.
b. `user_id`: Generated randomly based on the product's popularity, ensuring unique user IDs using a function.
c. `rating`: Generated by sampling from the rating distribution of the original dataset, ensuring a realistic distribution of ratings.
d. `Product title`: Taken from the `title_y` column, which represents the product title from the merged dataset.
e. `Categories`: Extracted directly from the `categories` column in the product dataset, ensuring the correct category is assigned to each product.
f. `timestamp`: Generated randomly by simulating a realistic date within the past year using a random offset.

Phase B: A prompt template was created to utilize the data points generated in Phase A, along with the Seed Dataset prepared in Step 1 (Seed Dataset Preparation), to generate outputs for the specified fields.

a. `title_x` (review title): Dynamically generated using a prompt to create a catchy review title based on the product name and rating.
b. `text` (review text): Dynamically generated using a prompt to create detailed review content based on the product category, title, and rating.
c. `helpful_vote`: Calculated based on the rating and length of the review, with randomness introduced to simulate realistic voting patterns.

Step 3: Testing and Validation

1. Textual Similarity Analysis:

- Measures how similar the synthetic reviews are to the original using methods like cosine similarity. Helps ensure the reviews aren't copied verbatim.

2. Lexical Diversity (Uniqueness in Word Usage):

- Measures the variety of vocabulary used in the synthetic dataset, ensuring it uses different words and phrases compared to the original dataset.

3. Distribution of Ratings and Other Features:

- Compares the distribution of features like ratings, review lengths, and helpful votes between the original and synthetic datasets to ensure pattern retention without replication.

4. N-gram Analysis (Word Patterns):

- Compares the frequency of word patterns (e.g., bigrams and trigrams) between the original and synthetic reviews, ensuring new phrases are introduced.

5. Perplexity (Model Confidence):

- Uses a language model to measure how "surprised" the model is by the synthetic reviews. A higher perplexity score indicates more novelty in the synthetic dataset.

6. Human Evaluation:

- A qualitative approach where human evaluators assess the realism, variety, and relevance of the synthetic reviews compared to the original dataset.

RESULTS

The final synthetic dataset closely resembled real Amazon reviews in structure, sentiment, and rating distribution. Reviews ranged from short, critical comments to longer, detailed reviews. Ratings followed a similar distribution to the original dataset, with an intentional imbalance where fewer 1-star reviews were generated.

CHALLENGES FACED

1. **Crafting Prompts:** Determining the parameters necessary to create prompts that would maintain realism was a key challenge.
2. **Addressing Repetition:** We needed to find solutions to overcome repetitive content in the generated reviews.
3. **Using Random Values:** We had to assess whether using random values for timestamps and helpful votes was feasible, as well as the potential consequences of doing so.
4. **Imbalance in Ratings:** The original dataset had fewer 1-star ratings, raising the question of whether the synthetic dataset should also reflect this imbalance.
5. **Formatting Issues:** Some of the earlier outputs had formatting problems, necessitating additional cleaning to ensure the reviews were presented in proper plain text.
6. **Balancing Review Length:** Striking the right balance between review length and content quality was challenging; longer reviews often sounded repetitive, while shorter reviews tended to lack detail.

FUTURE IMPROVEMENT

1. **Diversity in Review Content**: I plan to incorporate more personalized details and human-like anecdotes to make the reviews feel more authentic.
2. **Sentiment Analysis**: Fine-tuning the model to generate more nuanced sentiments in the reviews, especially for middle-range ratings, is another area for improvement.
3. **Aspect-Based Reviews**: In the future, reviews could focus on specific product aspects like price, packaging, or delivery, adding more depth to each review.
4. Leveraging open-source LLMs such as Meta's Llama 3.2 and Nvidia's NeMo 4-340B-Instruct can help decrease the expenses related to using GPT Turbo-3.
5. Comparison with alternative approaches such as:
    a. *https://sdv.dev/SDV/user_guides/single_table/ctgan.html*
    b. *https://mimesis.name/master/*
    c. *https://github.com/argilla-io/distilabel*

REFERENCES

1. Guo, X., Du, Z., Li, B., & Miao, C. Generating Synthetic Datasets for Few-shot Prompt Tuning. In *First Conference on Language Modeling* https://openreview.net/forum?id=Vd0KvChLXr#discussion
2. Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*. https://arxiv.org/abs/2406.15126
3. Kochanek, M., Cichecki, I., Kaszyca, O., Szydło, D., Madej, M., Jędrzejewski, D., ... & Kocoń, J. (2024). Improving Training Dataset Balance with ChatGPT Prompt Engineering. *Electronics*, *13*(12), 2255. https://www.mdpi.com/2079-9292/13/12/2255
4. Levi, E., Brosh, E., & Friedmann, M. (2024). Intent-based Prompt Calibration: Enhancing prompt optimization with synthetic boundary cases. *arXiv preprint arXiv:2402.03099*. https://arxiv.org/abs/2402.03099
5. Patil, R., Heston, T. F., & Bhuse, V. (2024). Prompt Engineering in Healthcare. *Electronics*, *13*(15), 2961. https://www.mdpi.com/2079-9292/13/15/2961