Title: Generating Synthetic Text Data using Large Language Models

Sahil Khandelwal

All the codes and the commented documentation can be found at:

https://github.com/sahilkhandelwal8/SyntheticDataWithLLM

Abstract

This assignment focuses on addressing the challenge of limited datasets available for training AI models across various tasks such as intent detection, slot filling, and recommendation systems. The objective is to generate synthetic datasets that closely resemble human-written text while enabling controlled variations in dialogue length, topic diversity, and language complexity. These synthetic datasets will be used for training, designing, and evaluating AI systems, offering a flexible solution to enhance model performance.

Understanding the Data

There are two given files – one with product data, the other with review data. These files can be merged based on **'parent_asin'** column, which refers to an umbrella product ID. Chemically identical products can have different individual IDs if they differ in colour, for example. But since their core properties are the same, they belong to one parent ID.

The product data is divided into subcategories, many of which have null values.

```
products['cat3'].isna().sum()
```

```
33814
```

```
products['cat4'].isna().sum()
```

```
187060
```

Presence of redundant columns

```
reviews['verified_purchase'].value_counts()
```

	count
verified_purchase	
True	16671

dtype: int64

This shows to be a redundant column.

From the plethora of available columns, thus a handful are selected.

The review written by a person depends on the **rating** they give to the product and the content of the review stems from a use case that can be inferred from the **product description**.

Initially, this data (rev1.csv) is used for trials.

	title	parent_asin	categories	rating	text
0	Bariatric Fusion Bariatric Multivitamin Soft C...	B0BKC2WYWB	['Health & Household'	5	Is a bariatric patient, my vitamins are really...
1	Bariatric Fusion Bariatric Multivitamin Soft C...	B0BKC2WYWB	['Health & Household'	1	Not only did it taste terrible with a horrible...
2	Bariatric Fusion Bariatric Multivitamin Soft C...	B0BKC2WYWB	['Health & Household'	5	Good flavor and no artificial sweeteners, whic...
3	Bariatric Fusion Bariatric Multivitamin Soft C...	B0BKC2WYWB	['Health & Household'	5	so Bariatric Fusion has multiple flavors and q...
4	Bariatric Fusion Bariatric Multivitamin Soft C...	B0BKC2WYWB	['Health & Household'	3	This looked like one of the best bariatric vit...

As the work progressed, further changes were made. This is what rev2.csv looks like –

	title	rating	text
0	Bariatric Fusion Bariatric Multivitamin Soft C...	5	Is a bariatric patient, my vitamins are really...
1	Bariatric Fusion Bariatric Multivitamin Soft C...	1	Not only did it taste terrible with a horrible...
2	Bariatric Fusion Bariatric Multivitamin Soft C...	5	Good flavor and no artificial sweeteners, whic...
3	Bariatric Fusion Bariatric Multivitamin Soft C...	5	so Bariatric Fusion has multiple flavors and q...
4	Bariatric Fusion Bariatric Multivitamin Soft C...	3	This looked like one of the best bariatric vit...

Further the column names are changed {'title':'product_name','text': 'product_review'}, for better data interpretation (rev3.csv).

Approaches Considered

- Base Instruct model (Implemented)
  Reasoning: Self-instruct models are better suited to create synthetic datasets.
- Trained Instruct model (Implemented)
- Gretel model (Implemented)

Reasoning: Gretel.ai is a company that provides a platform for creating synthetic data. The platform uses cutting-edge machine learning techniques to generate synthetic data that mimics real-world data, allowing organizations to train machine learning models without compromising data privacy or security.

- Conditional Text Generation (Discarded)
    - As the number of conditions increases, it becomes increasingly difficult for the model to generate data that satisfies everything
    - Specifically, in this case, the model can be made to generate reviews given two conditions [product_description, rating]
    - But enforcing additional constraints like dialogue length, topic diversity and language complexity hinders the data generation

- Retrieval-Augmented Generation (RAG) based approaches

Reasoning: I wondered if the model would work well if I was able to provide the model with few examples of what I want. For example, if I give existing reviews of a product with its [product_description, rating] and adequately prompt the agent, it should be able to do so with ease.

- Self – RAG models (Implemented)
- RAG with GPT (Implemented)
- RAG with Instruct model (Implemented)
- RAG with Gemini (Implemented)

Best Model: RAG with Gemini

Insights and Challenges from each trial

A sequential analysis of each trial.

Base Instruct model

Generates very generic ad-like reviews based on rating and description.

Input:

```
Generate a concise review strictly based on the Title, Categories, and Rating.Title: Liquid Multi
vitamin Nephronex by Llorens Pharmaceutical, 8 oz, B Complex Vitamins and Vitamin C Supplement, w
ith Biotin and Folic Acid, Fast Absorption, Adults and Children, Orange Flavor, Made in USA, Cate
gories: ['Health & Household', Rating: 5/5]
```

Output:

```
Review: Liquid Multivitamin Nephronex by Llorens Pharmaceutical is a top-rated supplement that st
ands out in the Health & Household category. This product offers a unique blend of B Complex Vita
mins, Vitamin C, Biotin, and Folic Acid, all essential nutrients for maintaining overall health.
The fast absorption feature ensures quick nutrient delivery to the body. Suitable for both adults
and children, it comes in an easy-to-use orange flavor. Made in the USA, Nephronex is a testament
to
```

Fine Tuned Instruct model

Displays unexpected behavior and unsatisfactory performance.

Input :

```
{'" Write a review for the product in 10 words. \nJudge how good or bad the product is based on t
he given rating. \nGive the review like a normal user considering cost, usecase,etc. '}
Title: 400lb / 180kg Digital Body Weight Bathroom Scale, High Precision Weighing Scale with Step-
On Technology, Tempered Right Angle Glass Balance Platform and Backlight Display, Categories: ['H
ealth & Household', Rating: 5
[INST]
```

Output:

```
choice for home use! High precision, easy to read backlit display, and sturdy design. Perfect for
tracking weight loss or maintenance."

Ground truth:
Great product great price.
```

Gretel model

The model shows great improvement and generates comprehensible and realistic reviews.

But the model is not able to capture the relationship between product,rating and review appropriately.

We see many contradictory examples like:

Product Name - *AnuMed Vitamin D3 Cream with Magnesium 10,000 IU | Healthy Skin Care & Face Cream | Maximum Calcium Absorption | Non-Greasy Moisturizer for Dry Skin - 3 Ounces*

Rating - *5*

Generated review - *These are nasty-tasting! Ewww! I threw them away*

Self -RAG model

Computationally expensive code – crashes even on GPU.

RAG with GPT2

Does not perform well. Ran through multiple iterations and could not get a satisfactory output.

The appropriate rows are retrieved perfectly. But the agent is not able to generate similar reviews. The prompt was also changed multiple times.

RAG with Instruct model

Similar issues as with GPT2. Not much improvement.

>> Since the retrieval worked well, I was confident that a better model can give desirable results.

RAG with Gemini model

The agent works really well. The reviews generated match the style, length variance, topic diversity of the sampled reviews.

Plus, the generation can be conditioned on the prompt.

Example : Prompt 1 – Write 5 new reviews for the product {product} .

Prompt 2 – Write 5 new reviews for the product {product} with rating {rating} .

Efficacy of results

Determining efficacy of synthetic text dataset is a challenging task. This is because we aim to quantitatively determine a qualitative result.

For this given task, these are the comparable quantifiers I have calculated for each (product,rating) pair

- **original average length vs synthetic average length**
- **original_avg_readability_score vs synthetic_avg_readability_score**
- **avg_semantic_similarity**

Readability score

The **Flesch Reading Ease** score is a quantitative measurement of how readable a piece of text is. The readability score that is assigned to a piece of writing is a number from 0 to 100. The higher the number is, the more readable the piece of text is.

The formula used is:

206.835 - 1.015 / (total number of words / total number of sentences) - 84.6 (total number of syllables / total number of words)

The score that a piece of text receives is calculated by using two key values: the text's average sentence length, and the average number of syllables in a word. The shorter these two values are, the easier the writing is to read. This is because long sentences that contain a large number of long, complex words require more effort to understand.

Semantic similarity

The reviews are turned into embeddings and cosine similarity is calculated.

Question

How do we ensure the synthetic dataset one generates is inspired from a source dataset but

not an exact replica?

>> A rich and diverse source dataset is crucial. Here, we also can tune the prompt and parameters like temperature, score_threshold, etc. to balance inspiration and replication.

Conclusion

The RAG with Gemini method tackles all our demands and allows for conditioned generation via prompt engineering.