Product Classification Using LLM

Introduction:

Overview:
In this project, we delve into the innovative use of Large Language Models (LLMs) for the challenging task of product classification. The project aims to harness the advanced capabilities of LLMs to categorize products accurately using limited and often incomplete data.

Objective:
Our main objective is to demonstrate how LLMs can effectively classify products despite challenges like missing data. The project focuses on overcoming these limitations through intelligent model training and prompt design.

Data Description:

The dataset utilized in this study presented significant challenges due to inconsistencies and gaps in crucial columns. Specifically, the 'product descriptions', 'categorization labels', and notably the often absent 'MARKETING_COPY' posed a challenge. These issues necessitated a tailored approach to ensure effective product classification.

Methodology:

1. LLM Classification Prompting:

The dataset consisted of several columns, notably: product_upc, product_name, item_brand_name, MARKETING_COPY, and product_image_url. Upon analysis, it was found that the MARKETING_COPY column had 999 null values, requiring special attention during the classification process.

```
df_product_classification.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2196 entries, 0 to 2195
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   product_upc        2196 non-null   int64 
 1   product_name       2196 non-null   object
 2   item_brand_name    2196 non-null   object
 3   MARKETING_COPY     1197 non-null   object
 4   product_image_url  2196 non-null   object
dtypes: int64(1), object(4)
memory usage: 85.9+ KB
```

```
# Checking NaN Values
df_product_classification.isnull().sum()
```

```
product_upc            0
product_name           0
item_brand_name        0
MARKETING_COPY       999
product_image_url      0
dtype: int64
```

Initially, the NULL values in the MARKETING_COPY column were neglected. A prompt was crafted to guide the LLM in identifying key features that suggest a product's category based on the available information: Product

Name, Item's Brand Name, and Marketing Copy. This step aimed to organize and document the model's classifications for further analysis and validation. The prompt that was used is shown below.

```
[64] prompt = "Identify key features that suggest a product's category based on the information provided. \
               The text has the Product Name, the Item's Brand Name, and a marketing copy of the item. \
               Utilizing all the information, identify the global product category of the product."
```

```
[68] df_final_data = pd.read_excel('/content/Final_data.xlsx')
```

```
df_final_data[['product_name','Response']][1235:1240]
```

1 to 5 of 5 entries Filter

index	product_name	Response
1235	Elderberry Dietary Supplement	omedicine technology provides potent antioxidants and immune-supporting benefits Answer: The product is a dietary supplement and its category is a health supplement.
1236	Strawberry Electrolyte Solution	electrolyte solution for dehydration relief Category: Beverages
1237	Moderate PADS, Lightly scented	otechnology designed to remove odor Category: Feminine Hygiene Products
1238	UDON PREMIUM NOODLE SOUP	-athin noodles, with a fresh, homemade flavor Category: Instant Noodle Soup
1239	Mtn Dew Code Red DEW With A Rush Of Cherry 2 L Bottle	This product is likely within the Beverages category, more specifically, Soft Drinks.

Show 25 per page

The above responses were recorded after running the prompts and loading it to the Excel sheet 'Final_data'. This step aimed to organize and document the model's classifications for further analysis and validation. However, while reviewing the model's performance, it was observed that the model exhibited unpredictable behavior, particularly when encountering products lacking marketing copy. These instances led to inconsistencies and inaccuracies in the generated classifications. This phenomenon is commonly referred to as "hallucinations," where the model produces responses that deviate significantly from the expected or accurate categorizations.
To address this challenge, a revised prompt was formulated to guide the model more effectively. The adjusted prompt instructed the model to identify product categories based on available attributes (Product Name and Item's Brand Name) even in cases where the marketing copy was missing.

```
import time
for index, row in df_product_classification.iterrows():
  df_product_classification['Response'][index] = get_response(row['Context'])
  time.sleep(5)
  if index == 3:
    break
```

```
df_product_classification[['product_name','Response']][:3]
```

	product_name	Response
0	e.l.f. 83314 Light Powder Blush Palette 1 ea	\n\nCosmetics Blush.
1	Amlactin Ultra Smoothing Intensely Hydrating C...	\n\nSkin Care
2	SOUR CREAM & ONION FLAVORED POPCORN	\n\nSnack Food

Accuracy Calculation:

When dealing with unlabeled data and employing an LLM classification model for labeling, assessing accuracy typically involves a method like sampling. One effective approach involves gathering categories for a subset of

brands using APIs such as UPC.org. By comparing these collected categories against the categories generated by the LLM classification model, we can evaluate the model's accuracy in assigning labels to the data.

Expanding on this process, it's crucial to ensure the subset of brands selected for category collection represents the diversity and range present in the overall dataset. Additionally, employing various statistical measures like precision, recall, and F1 score can provide a more comprehensive evaluation of the model's performance beyond simple accuracy.

2. Data Visualization:

```
# null count
summary = pd.DataFrame({'Total Rows': df_data_viz.shape[0], 'Null Rows': df_data_viz.isnull().sum()})
print(summary)
```

```
             Total Rows  Null Rows
UPC                4335       1737
PRODUCT_NAME       4335          0
BRAND_NAME         4335         12
ingredients        4335       2598
```

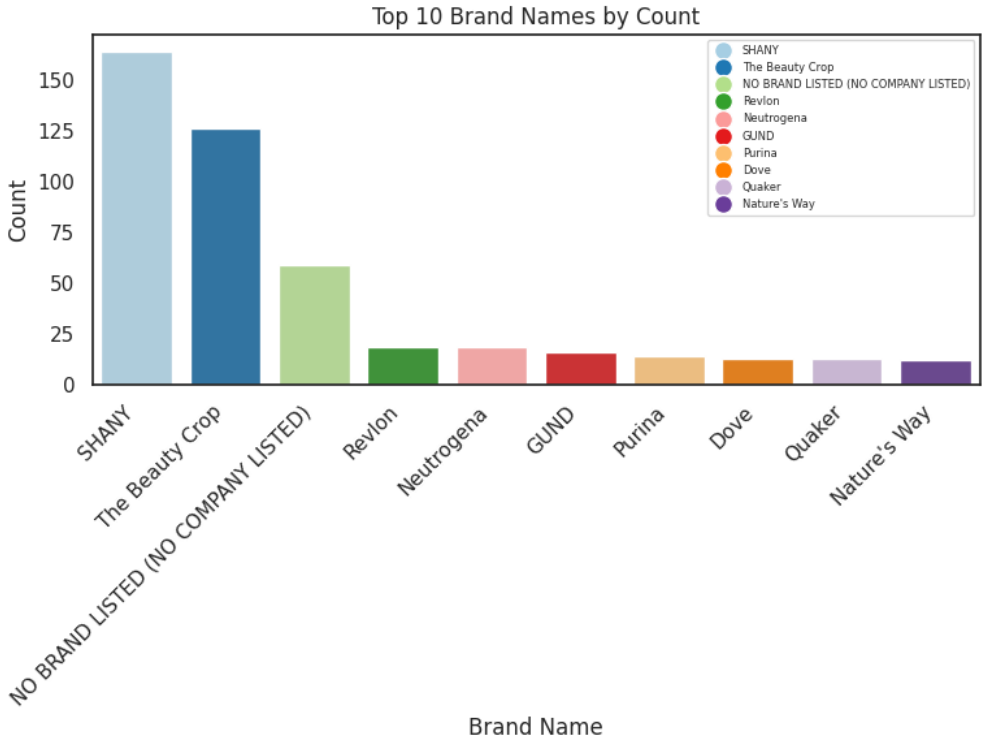
Considering the high proportion of missing data—40% in the UPC column and 60% in the ingredients column—it's advisable to retaining these columns to maintain data quality. However, as the BRAND_NAME column has minimal missing entries (less than 1%), dropping it would preserve valuable brand-related insights.

A word cloud was generated to depict the frequency and distribution of ingredients found across the products.

The word cloud provides a visual overview of the prevalent ingredients present in the products. Notably, recurring ingredients such as "Natural Ingredients," "Citric Acid," "Wheat Flour," "Iron Oxide," among others, emerged prominently. This observation suggests a prevailing trend within the market favoring natural components, common staple ingredients, and fortified elements.

The prominence of "Natural Ingredients" indicates a consumer inclination towards products perceived as more natural or organic. Similarly, the frequent occurrence of staple ingredients like "Wheat Flour" signifies the usage of fundamental components in various products, possibly indicating their widespread use across multiple categories. Moreover, the prevalence of "Iron Oxide" could signal a market emphasis on fortified products, likely aimed at addressing specific nutritional needs or enhancing product attributes.

Overall, these findings shed light on consumer preferences favoring healthier options and diverse nutritional attributes. Understanding these trends is pivotal for product development and marketing strategies aligned with consumer demands and preferences.

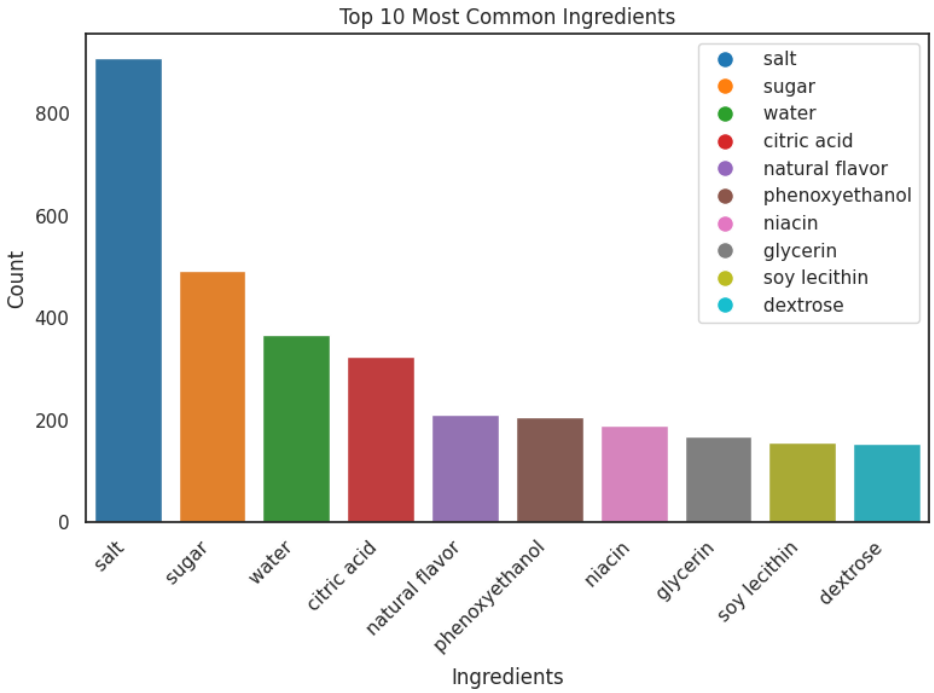


The chart shows that a variety of brands are represented in the top 10, including cosmetics brands (SHANY, The Beauty Crop, Revlon, Neutrogena), pet brands (GUND, Purina), and food brands (Dove, Quaker, Nature's Way). This suggests that the data is representative of a wide range of products and industries. SHANY and The Beauty Crop are the two most common brand names in the data, each with over 150 occurrences. It is interesting to note that two of the top three brands are cosmetics brands, suggesting that cosmetics are a popular category of products.

Overall, the chart shows that the top 10 brand names by count in the United States are all consumer goods brands. This suggests that consumers are more likely to purchase products from brands that they are familiar with and trust. The top 5 brand names on the list are all beauty brands (SHANY, The Beauty Crop, Revlon, Neutrogena). This suggests that the beauty industry is a major driver of consumer spending in the United States.

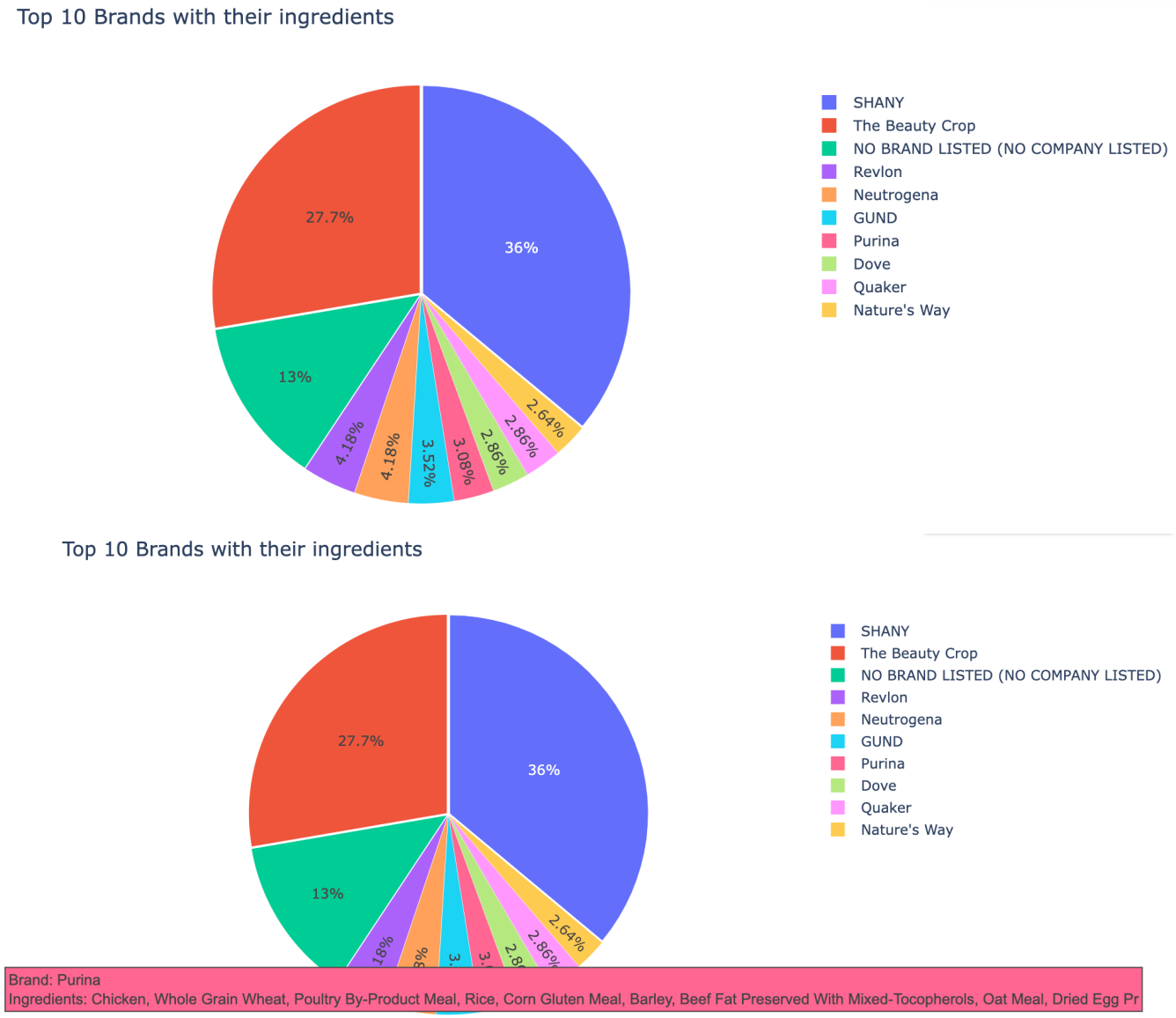
The other 5 brand names on the list are a mix of and food and beverage brands (Dove, Quaker, Nature's Way) and pet brands (GUND, Purina). This suggests that consumers are also interested in purchasing high-quality products from these categories from brands that they trust.

Interestingly, there is no single brand that dominates the list. This suggests that there is a healthy level of competition in the consumer goods industry in the United States.



The most common ingredient is salt, followed by sugar, water, citric acid, natural flavor, phenoxyethanol, niacin, glycerin, soy lecithin, and dextrose. It is interesting to note that the top 10 most common ingredients are all relatively simple substances. This suggests that many of the products that we use in our everyday lives are made up of a relatively small number of basic ingredients. The top 5 most common ingredients are all food ingredients. This suggests that food is the largest category of products that use these ingredients.

The next 5 most common ingredients are all used in a variety of different products, including food, cosmetics, and personal care products. This suggests that these ingredients are very versatile and can be used in a wide range of applications.



The largest slice of the pie is for Purina, with 36%. This is likely because Purina is a popular brand of pet food. The next largest slice is for SHANY, with 27.76%. SHANY is a cosmetics brand that is known for its affordable and high-quality products. Additionally, the fact that all of the brands on the list are known for their natural ingredients suggests that there is a growing demand for these types of products.

Similarity Analysis using BERT:

To further explore the dataset and understand the relationships between products based on textual data within the ingredients, advanced models such as BERT (Bidirectional Encoder Representations from Transformers) and SentenceTransformer were employed.

Utilizing BERT's tokenizer and the SentenceTransformer library, numerical embeddings of fixed sizes were derived from the textual data present in the ingredients column.

	index	UPC	PRODUCT_NAME	BRAND_NAME	ingredients	embeddings
0	0	NaN	Tillamook Farmstyle Thick Cut Medium Cheddar S...	Tillamook	Medium Cheddar Cheese (Cultured Milk, Salt, En...	[-0.2700659, -0.4764282, 0.59425163, 1.1468518...
1	1	NaN	Nice N Easy Permanent Creme Color 1 ea	Nice N Easy	Gray Retexturizing Pre-Treatment: Water, Propy...	[0.23007461, -0.296552, 0.8529975, 0.64112484,...
2	3	NaN	Skin Effects Acne Spot Treatment 0.25 oz	Skin Effects	Water (Aqua), Squalane, Butylene Glycol, Allyl...	[-0.30950654, -0.18036936, 0.108600795, 0.7994...
3	5	NaN	Pride Lip Balm & Scrub Set	The Beauty Crop	LIP BALM: ETHYLHEXYL PALMITATE, ISONONY ISONON...	[0.44993877, -0.24951757, 0.50882876, -0.13076...
4	11	NaN	Fierce & Flawless All-in-One Compact with 34 C...	SHANY	EYE SHADOW INGREDIENTS: Mica, Dimethicone, Tal...	[0.18113355, -0.30725268, -0.23870845, 0.74527...

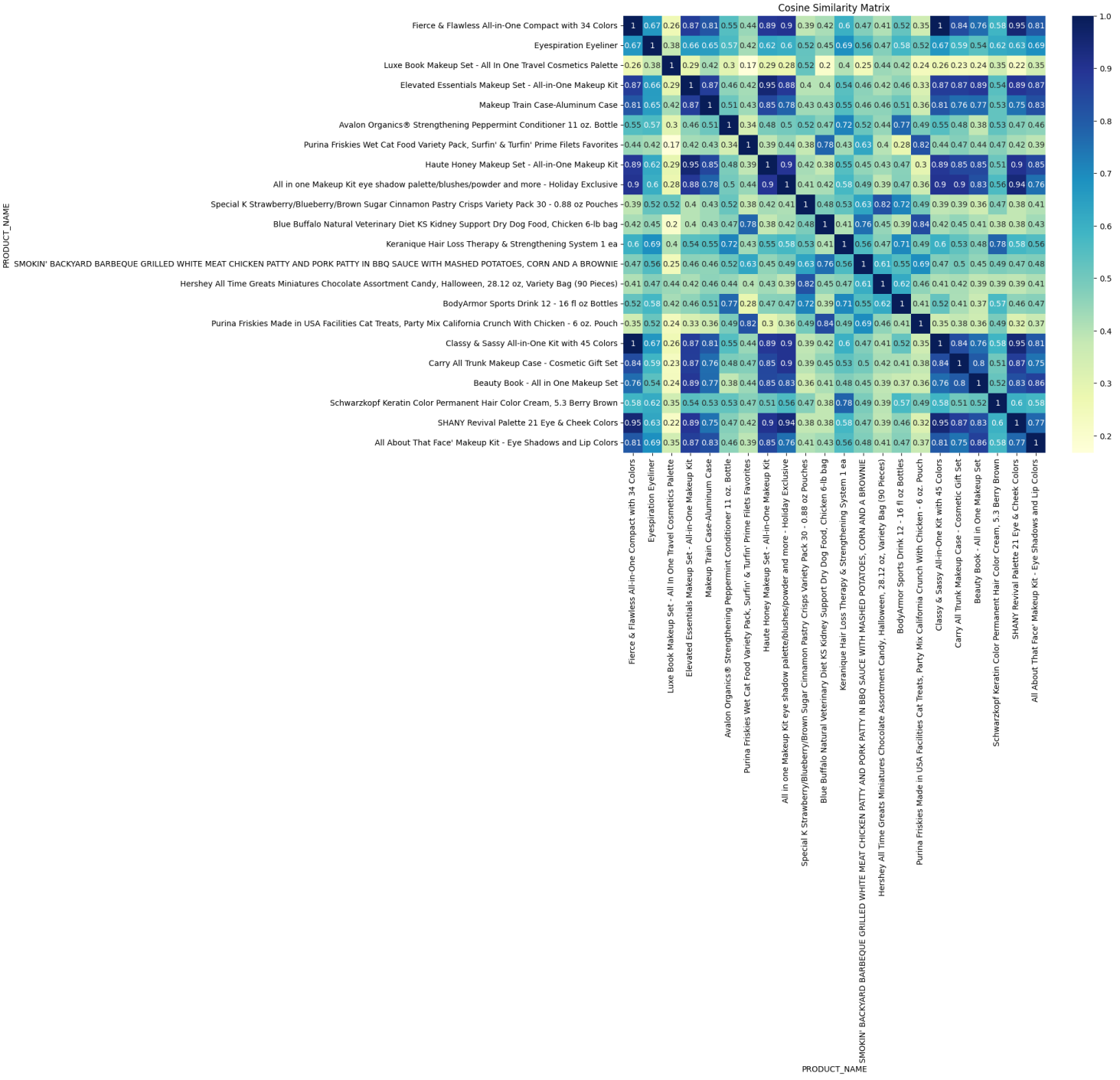
These embeddings facilitated the calculation of cosine similarity scores for each product's ingredients. The similarity analysis produced compelling results, demonstrating the relationships between products based on ingredient similarities.

For instance, consider the example for the product "XTRA CHEDDAR BAKED SNACK CRACKERS." The analysis generated the top five most similar products along with their respective similarity scores

Five Most similar products for XTRA CHEDDAR BAKED SNACK CRACKERS

	Similar Products	Similarity Score
0	CHEDDAR JALAPENO MEGA BITES BAKED SNACK CRACKERS	0.935477
1	Goldfish Disney Princess Cheddar Baked Snack C...	0.921035
2	BAKED SNACK CRACKERS	0.886940
3	Goldfish Family Size Colors Colors Cheddar Bak...	0.873913
4	STUFFED PRETZEL BREAD SANDWICH	0.853017

From the cosine similarity score, it can be noticed that the ‘Cheddar Jalapeno Mega Bites Snack Crackers’ was the most similar to our Traget Product which is also a Snack Cracker.



Let’s take a closer look at this plot for our analysis:

Based on ingredients, the SHANY and Fierce and Flawless are very similar. The remarkably high similarity score of 0.95 suggests a significant overlap in the ingredients or formulation between the products SHANY and Fierce and Flawless. Such a substantial similarity might indicate these products belonging to the same product line, sharing core ingredients, or targeting similar consumer preferences and needs. These products might be positioned as alternatives to each other, catering to diverse consumer preferences within a specific category.

On the contrary, if a food product and cosmetics share high similarities, we can predict that if a customer likes the food, they will also like the cosmetic and potentially recommend it to them.

3. Generating a Marketing Copy

The process involved creating a specific marketing prompt that outlined the requirements for generating a persuasive marketing copy emphasizing the health-related benefits of a product.

A Python function, get_response(context), was developed to interact with OpenAI's GPT-3 API (openai.completions.create). This function incorporated parameters such as the model type (text-davinci-003), a temperature setting of 0 for precision, token limits, and the number of desired responses.

```
print(marketing_prompt)
```

```
generate a brief, coherent marketing copy for a set of products, using the product name and item brand name given below.
The model should output a short, persuasive marketing copy that highlights the unique feature of the product that makes
someone healthier. And remember, your marketing copy must be legal!
```

```
def get_response(context):
  response = openai.completions.create(
      model="text-davinci-003",
      temperature = 0,
      prompt= marketing_prompt + context,
      max_tokens=150,
      n=1,
      stop=None,
  )
  classified_category = response.choices[0].text
  return classified_category
```

```
marketing_context = 'Product Name - Tillamook Farmstyle Thick Cut Medium Cheddar Shredded Cheese 8 oz \
and Brand Name - Tillamook '
```

```
print(get_response(marketing_context))
```

```
Tillamook Farmstyle Thick Cut Medium Cheddar Shredded Cheese 8 oz is the perfect way to add flavor and nutrition to your
meals. Our cheese is made with only the finest ingredients, and is aged to perfection for a rich, creamy flavor. Our
medium cheddar is a great source of calcium and protein, making it a healthier choice for your family. Plus, it's pre-
shredded, so you can easily add it to your favorite dishes. Enjoy the delicious taste of Tillamook Farmstyle Thick Cut
Medium Cheddar Shredded Cheese 8 oz today!
```

Providing specific product and brand details as the context, the function generated a persuasive marketing copy. For instance, the product ‘Tillamook Farmstyle Thick Cut Medium Cheddar Shredded Cheese 8 oz’ was selected. The resulting output emphasized the product's nutritional value, superior taste, use of quality ingredients, and its convenience in meal preparation. It highlighted the health benefits such as being a rich source of calcium and protein, appealing to families seeking healthier options.

Ultimately, the generated marketing copy aligned with the outlined criteria, effectively showcasing the unique selling points of the product while emphasizing its health benefits and delicious taste.

Conclusion:

This project successfully demonstrated the potential of Large Language Models in product classification, even in scenarios with limited data availability. The insights gained open avenues for future research, especially in improving LLM prompt design and exploring other applications of LLMs in data analysis and classification.