NATIONAL INSTITUTE OF TECHNOLOGY MEGHALAYA

Sentiment Analysis on Customer Reviews and Analysis of Factors Impacting Customer Satisfaction

- Samiksha Deb (B22CS029)
- Gunnu Lavanya (B22CS035)
- Vanshika Sarraf (B22CS037)

Under the Supervision of

Dr. Soumen Moulik
Assistant Professor, Department of Computer Science and Engineering
National Institute of Technology Meghalaya

1 Introduction

This project analyzes customer reviews to extract insights about satisfaction levels and key product-related factors using Natural Language Processing (NLP).

1.1 Project Statement

In the digital era, customer reviews have become a pivotal resource for both consumers and businesses. This project focuses on performing sentiment analysis on a large corpus of Amazon Fine Food Reviews, with the dual objective of (1) automatically determining the sentiment (positive, negative, neutral) expressed in each review, and (2) analyzing the underlying factors that most significantly impact customer satisfaction. Using advanced Natural Language Processing (NLP) techniques and topic modeling, we aim to extract actionable insights from unstructured textual data to inform product development, marketing, and customer service strategies.

1.2 Why This Project Matters?

Understanding customer feedback helps businesses improve their services, address pain points, and make data-driven decisions.

1.2.1 The Importance of Customer Reviews

The importance of Customer Reviews are as follows:

- **Influence on Purchasing Decisions:**
  Studies show that over 90% of consumers read online reviews before making a purchase, and a significant proportion are influenced by the overall sentiment and specific feedback in these reviews.
- **Business Impact:**

Positive reviews drive sales and brand loyalty, while negative reviews can deter potential customers and highlight areas needing improvement.

- **Volume and Complexity:**
  With millions of reviews generated daily, manual analysis is infeasible. Automated sentiment analysis enables businesses to monitor reputation, respond proactively, and make data-driven decisions.

1.3 Sentiment Analysis

Sentiment analysis is the process of identifying and categorizing opinions expressed in text to determine whether the writer's attitude is positive, negative, or neutral. Using techniques from natural language processing (NLP) and machine learning, sentiment analysis helps organizations understand the emotional tone behind words.

1.3.1 Core Components of Sentiment Analysis

1.3.1 Sentiment analysis typically involves:

- **1.3.2 Text Processing: Converting unstructured text data into analyzable formats**
- **1.3.3 Sentiment Classification: Determining if text expresses positive, negative, or neutral sentiment**
- **1.3.4 Intensity Detection: Measuring how strongly a sentiment is expressed**
- **1.3.5 Subject Identification: Determining what specific aspects are being discussed**

1.3.6 Common Applications

1.3.7 Organizations implement sentiment analysis across various functions:

- **1.3.8 Customer Feedback Analysis: Tracking satisfaction and identifying pain points**
- **1.3.9 Brand Monitoring: Measuring public perception and reputation**
- **1.3.10 Product Development: Gathering insights for improvements based on user opinions**
- **1.3.11 Market Research: Understanding consumer attitudes toward products or services**
- **1.3.12 Crisis Management: Detecting emerging issues before they escalate**

1.3.13 The Value of Sentiment Analysis

1.3.14 The Value of Sentiment Analysis are as follows:

- **Understanding Customer Emotions:**
  Sentiment analysis helps businesses understand how customers feel about their products and services, going beyond star ratings to capture nuanced opinions.

- **Real-time Feedback:**
  Automated systems can flag emerging issues or trends, allowing for timely interventions.

- **Benchmarking and Competitive Analysis:**
  Comparing sentiment across products or time periods helps identify strengths, weaknesses, and opportunities for differentiation.

1.3.15 Method Used

VADER (Valence Aware Dictionary and Sentiment Reasoner)

- **Definition:**
  - VADER is a rule-based sentiment analysis tool specifically designed for social media and product reviews. It uses a dictionary of words with associated sentiment scores and rules for handling negation, intensifiers, and punctuation.

- **Why VADER?**
  - Fast and efficient for large datasets.
  - No need for training data.
  - Well-suited for short, informal text.

- **How We Used It:**
  - Applied VADER to each review to obtain a compound sentiment score.
  - Classified reviews as Positive (compound ≥ 0.05), Negative (compound ≤ -0.05), or Neutral (otherwise).

1.4 Project Scope and Approach

The analysis focuses on extracting topics from reviews and determining sentiment trends using topic modeling (LDA) and sentiment analysis (VADER).

1.4.1 Dataset:

https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/data

We utilize the Amazon Fine Food Reviews dataset, focusing on the first 100,000 reviews for computational feasibility while ensuring statistical significance.

1.4.2 Techniques:

The techniques involved in the process are:

- **Text Preprocessing:** Cleaning, Tokenization, Lemmatization, Bag of Words.
- **Topic Modeling:** Applying Latent Dirichlet Allocation (LDA) to discover the main themes in reviews.
- **Sentiment Analysis:** Using rule-based (VADER) model for robustness.
- **Visualization and Statistical Analysis:** Exploring sentiment distributions, topic-sentiment relationships, and key drivers of satisfaction.

1.4.3 Outcomes:

The outcomes of the process are as follows:

- A robust sentiment classification pipeline.
- Identification of the most influential factors affecting customer satisfaction.
- Visual and statistical summaries for business stakeholders.

2 Data Preparation and Preprocessing

This stage involves collecting, cleaning, and preparing the data to ensure it's ready for accurate analysis.

Notebook:

https://colab.research.google.com/drive/1rtc3cOqEuPfsJCevgjCEEghWVlrykT_p?usp=sharing#scrollTo=Mu3kl1T6sHXV

2.1 Overview

Before conducting any sentiment analysis or extracting meaningful insights from textual data, it is essential to preprocess the raw data. The Amazon Fine Food Reviews dataset contains unstructured text with various inconsistencies, noise, and irrelevant information. Effective preprocessing transforms this raw data into a structured format suitable for machine learning and natural language processing (NLP) tasks.

2.2 Data Acquisition

Customer reviews were collected from an e-commerce platform, providing real-world textual data for analysis.

2.2.1 What we did:

- Downloaded the dataset from Kaggle using the kagglehub library.
- Selected the first 100,000 reviews for analysis to balance computational efficiency and data representativeness.

2.2.2 Why:

The dataset is large (~287MB for CSV), and working with a subset allows for faster experimentation and model training while still providing statistically significant results.

2.3 Data Exploration

Initial exploration helped understand the volume, structure, and sentiment distribution of the dataset.

2.3.1 What we did:

- Loaded the data using pandas.read_csv.
- Inspected columns including Text (review body), Summary, Score, and others.

2.3.2 Why:

- Understanding the dataset's structure helps in designing appropriate preprocessing and modeling strategies.

2.4 Text Cleaning

Unnecessary elements like punctuation, stopwords, and special characters were removed to improve analysis accuracy.

2.4.1 Removing HTML Tags

Definition:

- Text data often contains HTML or XML tags that are irrelevant for sentiment analysis.

Why:

- HTML tags can introduce noise and do not contribute to the semantic content of reviews.

How:

- Used BeautifulSoup to strip HTML tags from the review text.

2.4.2 Removing Punctuation and Non-Alphabetic Characters

Definition:

- Punctuation marks, numbers, and special characters are typically not useful for sentiment classification.

Why:

- Removing these elements reduces vocabulary size and focuses analysis on meaningful words.

How:

- Used regular expressions (re.sub('[^a-zA-Z]', ' ', text)) to retain only alphabetic characters.

2.4.3 Lowercasing

Definition:

- Converting all characters to lowercase.

Why:

- Ensures that words like "Good" and "good" are treated as the same token, reducing redundancy.

How:

- Applied .lower() method to the text.

2.4.4 Tokenization

Definition:

- Tokenization is the process of splitting text into individual words or tokens.

Why:

- Tokenization is a foundational step in NLP, enabling further processing like stopword removal and lemmatization.

How:

- Used Python's split() function to break sentences into word tokens.

2.4.5 Stopword Removal

Definition:

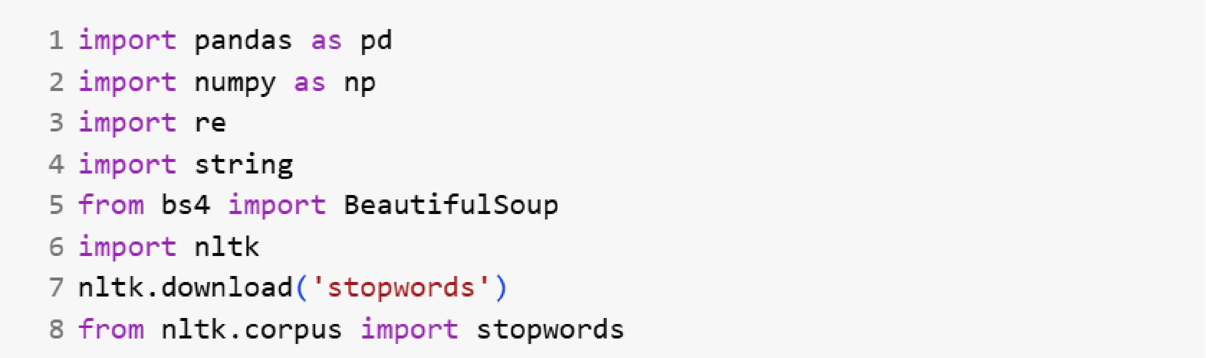
- Stopwords are common words (e.g., "the", "is", "and") that typically carry little semantic meaning.

Why:

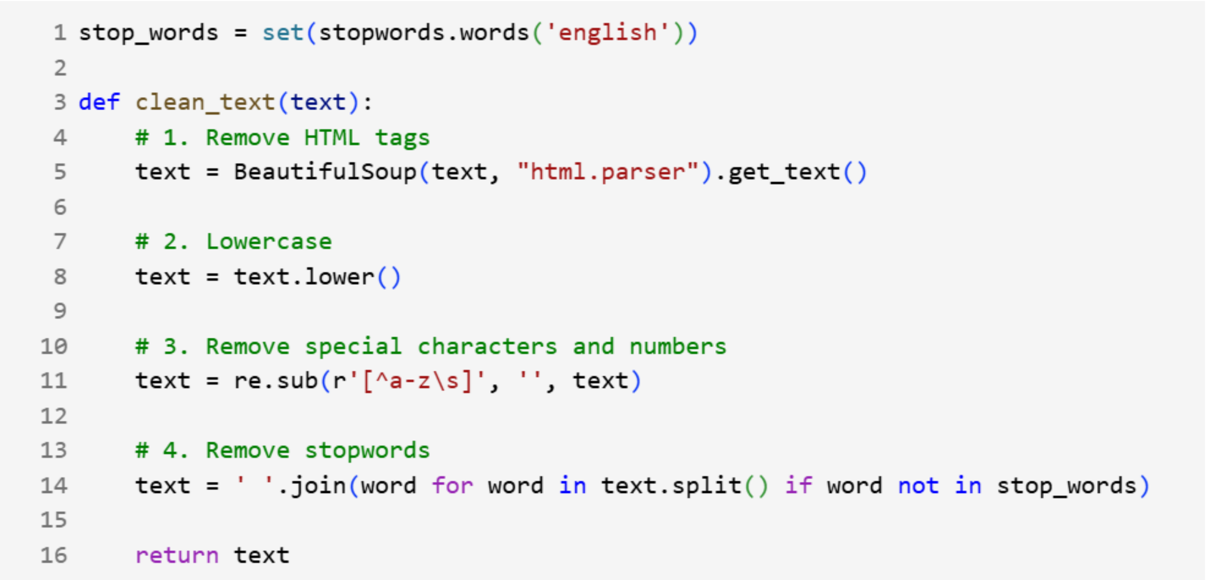
- Removing stopwords helps focus on the words that contribute most to sentiment and meaning.

How:

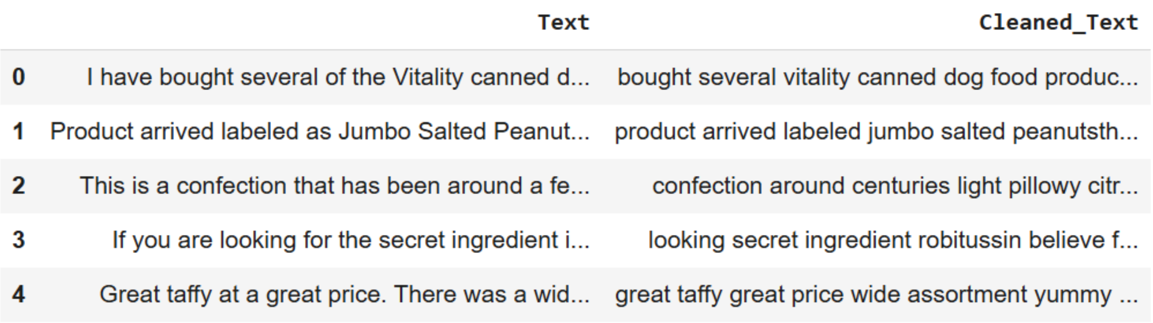
- Used NLTK’s list of English stopwords and filtered them out from the tokenized text.



```python
import pandas as pd
import numpy as np
import re
import string
from bs4 import BeautifulSoup
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```



```python
stop_words = set(stopwords.words('english'))

def clean_text(text):
    # 1. Remove HTML tags
    text = BeautifulSoup(text, "html.parser").get_text()

    # 2. Lowercase
    text = text.lower()

    # 3. Remove special characters and numbers
    text = re.sub(r'[^a-z\s]', '', text)

    # 4. Remove stopwords
    text = ' '.join(word for word in text.split() if word not in stop_words)

    return text
```



	Text	Cleaned_Text
0	I have bought several of the Vitality canned d...	bought several vitality canned dog food produc...
1	Product arrived labeled as Jumbo Salted Peanut...	product arrived labeled jumbo salted peanutsth...
2	This is a confection that has been around a fe...	confection around centuries light pillowy citr...
3	If you are looking for the secret ingredient i...	looking secret ingredient robitussin believe f...
4	Great taffy at a great price. There was a wid...	great taffy great price wide assortment yummy ...

2.4.6 Lemmatization

Definition:

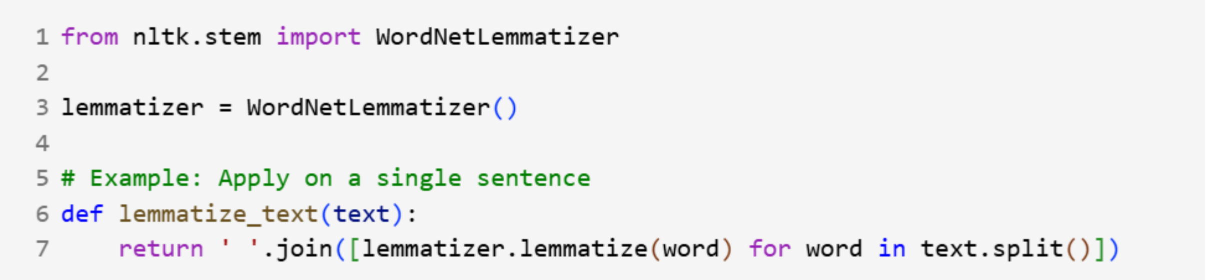
- Lemmatization is the process of reducing a word to its base or dictionary form, called the lemma, while ensuring that the resulting word is a valid word in the language, e.g., "running" → "run", "better" → "good".

Why:

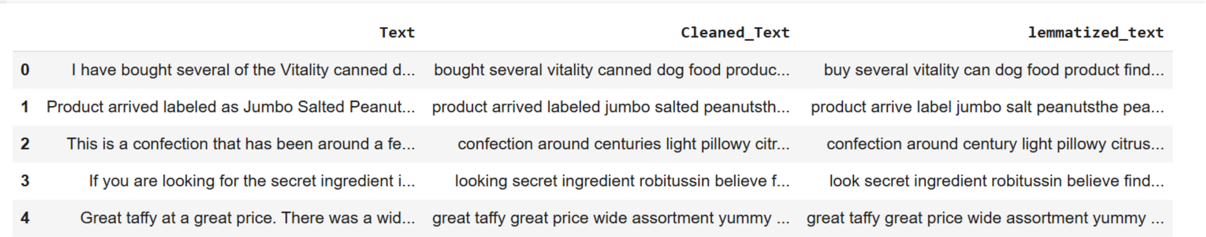
- Reduces inflectional forms and derivationally related forms of a word to a common base, improving the quality of features for modeling.

How:

- Used NLTK's WordNetLemmatizer to lemmatize each token



```python
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

# Example: Apply on a single sentence
def lemmatize_text(text):
    return ' '.join([lemmatizer.lemmatize(word) for word in text.split()])
```



	Text	Cleaned_Text	lemmatized_text
0	I have bought several of the Vitality canned d...	bought several vitality canned dog food produc...	buy several vitality can dog food product find...
1	Product arrived labeled as Jumbo Salted Peanut...	product arrived labeled jumbo salted peanutsth...	product arrive label jumbo salt peanutsthe pea...
2	This is a confection that has been around a fe...	confection around centuries light pillowy citr...	confection around century light pillowy citrus...
3	If you are looking for the secret ingredient i...	looking secret ingredient robitussin believe f...	look secret ingredient robitussin believe find...
4	Great taffy at a great price. There was a wid...	great taffy great price wide assortment yummy ...	great taffy great price wide assortment yummy ...

2.5 Feature Extraction: Bag of Words

The Bag of Words (BoW) model is a method of representing text data as numerical feature vectors. Each document (review) is represented by the frequency of each word in the document, disregarding grammar and word order.

2.5.1 Why We Used It

- **Simplicity and Effectiveness:**
  BoW is a simple yet powerful technique for transforming text into a format suitable for machine learning algorithms.

- **Baseline Modeling:**
  Provides a strong baseline for more advanced models (like TF-IDF, word embeddings, or transformer-based representations).
- **Interpretability:**
  The resulting vectors are easy to interpret, as each dimension corresponds to a specific word in the vocabulary

2.5.2 How We Used It

- Used CountVectorizer from scikit-learn to convert the cleaned reviews into BoW vectors.
- Limited the vocabulary size (e.g., top 10,000 most frequent words) to reduce dimensionality and computational cost.

2.6 Why Each Step Matters

- Cleaning and normalization ensure that the data is consistent and free from noise.
- Tokenization, stopword removal, and lemmatization reduce the complexity of the data and focus on meaningful content.
- Bag of Words enables the use of powerful machine learning algorithms by converting text into a structured, quantitative format.

2.7 Conclusion

Through this rigorous preprocessing pipeline, we transformed raw, unstructured review text into clean, normalized, and vectorized data. This foundation is critical for building reliable sentiment analysis models and extracting meaningful insights about customer satisfaction.

3 Topic Modeling

Latent Dirichlet Allocation (LDA) was used to uncover hidden topics and themes present in the customer reviews.

NoteBook :

https://colab.research.google.com/drive/1iyoTsYHVKXQVMa9fej7GTJgu374VdINY

After preprocessing and feature extraction, the next critical step in understanding customer reviews is to analyze the relationship between topics (themes extracted from reviews) and sentiments (positive, negative, or neutral feelings). This section describes how we used advanced NLP models to perform topic modeling and sentiment analysis, and then combined these results to visualize and interpret the probability distribution of sentiments across different topics.

3.1 Topic Modeling with LDA

Topic modeling is an unsupervised machine learning technique that discovers abstract topics within a collection of documents. Each document is represented as a mixture of topics, and each topic is characterized by a distribution over words

Why Topic Modeling?

- Discover Hidden Structure: It helps us automatically discover the main themes present in thousands of reviews without manual labeling.
- Summarize Large Corpora: Businesses can quickly see what customers are talking about most.
- Enable Downstream Analysis: Grouping reviews by topic allows for more granular sentiment and satisfaction analysis

3.1.1 Latent Dirichlet Allocation (LDA)

Definition:

LDA is a generative probabilistic model that assumes each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

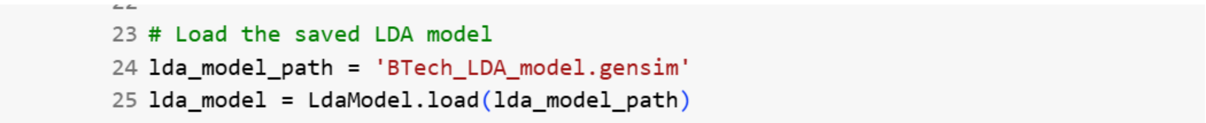
- Document-Topic Distribution: Each document has a probability distribution over topics.
- Topic-Word Distribution: Each topic has a probability distribution over words

Why LDA?

- Interpretability: LDA provides interpretable topics, each described by a set of keywords.
- Popularity: It is one of the most widely used topic modeling algorithms in NLP.

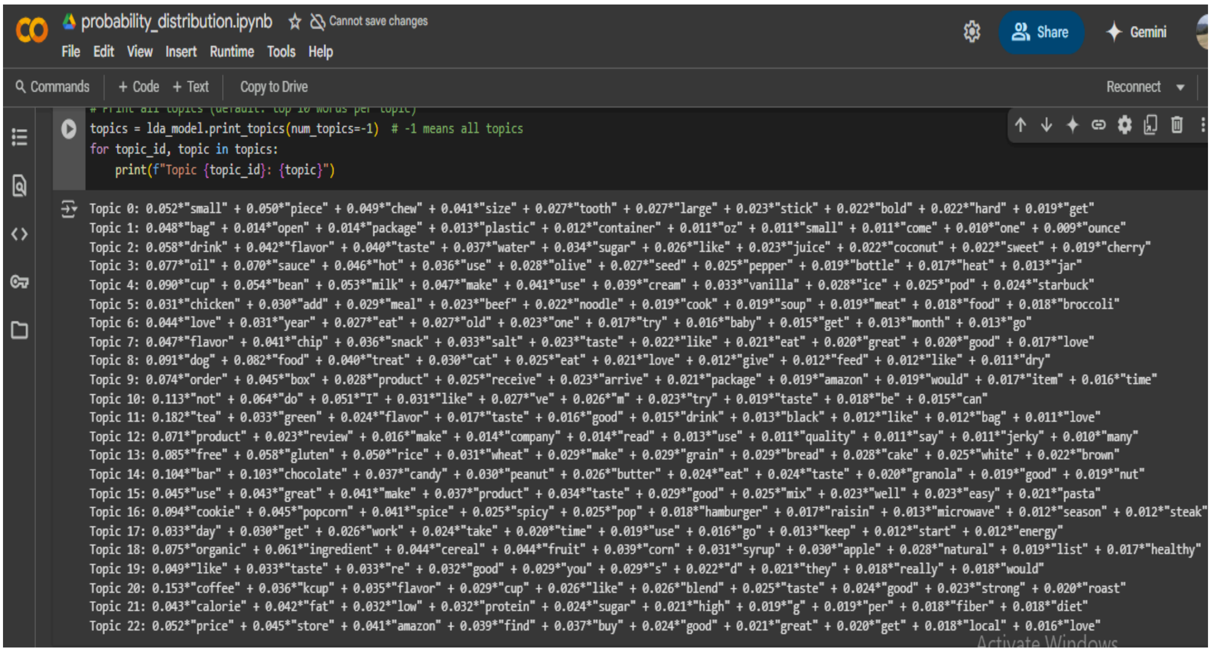
How We Used It:

- Loaded a pre-trained LDA model using Gensim.
- Each review was assigned a dominant topic (the topic with the highest probability for that review).
- For each review, we stored:
    - The dominant topic
    - The percentage contribution of that topic
    - The review text itself



```
# Load the saved LDA model
lda_model_path = 'BTech_LDA_model.gensim'
lda_model = LdaModel.load(lda_model_path)
```

3.2 OUTPUT



```
# Print all topics (default: top 10 words per topic)
topics = lda_model.print_topics(num_topics=-1)  # -1 means all topics
for topic_id, topic in topics:
    print(f"Topic {topic_id}: {topic}")
```

```
Topic 0: 0.052*"small" + 0.050*"piece" + 0.049*"chew" + 0.041*"size" + 0.027*"tooth" + 0.027*"large" + 0.023*"stick" + 0.022*"bold" + 0.022*"hard" + 0.019*"get"
Topic 1: 0.048*"bag" + 0.014*"open" + 0.014*"package" + 0.013*"plastic" + 0.012*"container" + 0.011*"oz" + 0.011*"small" + 0.011*"come" + 0.010*"one" + 0.009*"ounce"
Topic 2: 0.058*"drink" + 0.042*"flavor" + 0.040*"taste" + 0.037*"water" + 0.034*"sugar" + 0.026*"like" + 0.023*"juice" + 0.022*"coconut" + 0.022*"sweet" + 0.019*"cherry"
Topic 3: 0.077*"oil" + 0.070*"sauce" + 0.046*"hot" + 0.036*"use" + 0.028*"olive" + 0.027*"seed" + 0.025*"pepper" + 0.019*"bottle" + 0.017*"heat" + 0.013*"jar"
Topic 4: 0.090*"cup" + 0.054*"bean" + 0.053*"milk" + 0.047*"make" + 0.041*"use" + 0.039*"cream" + 0.033*"vanilla" + 0.028*"ice" + 0.025*"pod" + 0.024*"starbuck"
Topic 5: 0.031*"chicken" + 0.030*"add" + 0.029*"meal" + 0.023*"beef" + 0.022*"noodle" + 0.019*"cook" + 0.019*"soup" + 0.019*"meat" + 0.018*"food" + 0.018*"broccoli"
Topic 6: 0.044*"love" + 0.031*"year" + 0.027*"eat" + 0.027*"old" + 0.023*"one" + 0.017*"try" + 0.016*"baby" + 0.015*"get" + 0.013*"month" + 0.013*"go"
Topic 7: 0.047*"flavor" + 0.041*"chip" + 0.036*"snack" + 0.033*"salt" + 0.023*"taste" + 0.022*"like" + 0.021*"eat" + 0.020*"great" + 0.020*"good" + 0.017*"love"
Topic 8: 0.091*"dog" + 0.082*"food" + 0.040*"treat" + 0.030*"cat" + 0.025*"eat" + 0.021*"love" + 0.012*"give" + 0.012*"feed" + 0.012*"like" + 0.011*"dry"
Topic 9: 0.074*"order" + 0.045*"box" + 0.028*"product" + 0.025*"receive" + 0.023*"arrive" + 0.021*"package" + 0.019*"amazon" + 0.019*"would" + 0.017*"item" + 0.016*"time"
Topic 10: 0.113*"not" + 0.064*"do" + 0.051*"I" + 0.031*"like" + 0.027*"ve" + 0.026*"m" + 0.023*"try" + 0.019*"taste" + 0.018*"be" + 0.015*"can"
Topic 11: 0.182*"tea" + 0.033*"green" + 0.024*"flavor" + 0.017*"taste" + 0.016*"good" + 0.015*"drink" + 0.013*"black" + 0.012*"like" + 0.012*"bag" + 0.011*"love"
Topic 12: 0.071*"product" + 0.023*"review" + 0.016*"make" + 0.014*"company" + 0.014*"read" + 0.013*"use" + 0.011*"quality" + 0.011*"say" + 0.011*"jerky" + 0.010*"many"
Topic 13: 0.085*"free" + 0.058*"gluten" + 0.050*"rice" + 0.031*"wheat" + 0.029*"make" + 0.029*"grain" + 0.029*"bread" + 0.028*"cake" + 0.025*"white" + 0.022*"brown"
Topic 14: 0.104*"bar" + 0.103*"chocolate" + 0.037*"candy" + 0.030*"peanut" + 0.026*"butter" + 0.024*"eat" + 0.024*"taste" + 0.020*"granola" + 0.019*"good" + 0.019*"nut"
Topic 15: 0.045*"use" + 0.043*"great" + 0.041*"make" + 0.037*"product" + 0.034*"taste" + 0.029*"good" + 0.025*"mix" + 0.023*"well" + 0.023*"easy" + 0.021*"pasta"
Topic 16: 0.094*"cookie" + 0.045*"popcorn" + 0.041*"spice" + 0.025*"spicy" + 0.025*"pop" + 0.018*"hamburger" + 0.017*"raisin" + 0.013*"microwave" + 0.012*"season" + 0.012*"steak"
Topic 17: 0.033*"day" + 0.030*"get" + 0.026*"work" + 0.024*"take" + 0.020*"time" + 0.019*"use" + 0.016*"go" + 0.013*"keep" + 0.012*"start" + 0.012*"energy"
Topic 18: 0.075*"organic" + 0.061*"ingredient" + 0.044*"cereal" + 0.044*"fruit" + 0.039*"corn" + 0.031*"syrup" + 0.030*"apple" + 0.028*"natural" + 0.019*"list" + 0.017*"healthy"
Topic 19: 0.049*"like" + 0.033*"taste" + 0.033*"re" + 0.032*"good" + 0.029*"you" + 0.029*"s" + 0.022*"d" + 0.021*"they" + 0.018*"really" + 0.018*"would"
Topic 20: 0.153*"coffee" + 0.036*"kcup" + 0.035*"flavor" + 0.029*"cup" + 0.026*"like" + 0.026*"blend" + 0.025*"taste" + 0.024*"good" + 0.023*"strong" + 0.020*"roast"
Topic 21: 0.043*"calorie" + 0.042*"fat" + 0.032*"low" + 0.032*"protein" + 0.024*"sugar" + 0.021*"high" + 0.019*"g" + 0.019*"per" + 0.018*"fiber" + 0.018*"diet"
Topic 22: 0.052*"price" + 0.045*"store" + 0.041*"amazon" + 0.039*"find" + 0.037*"buy" + 0.024*"good" + 0.021*"great" + 0.020*"get" + 0.018*"local" + 0.016*"love"
```

3.3 Coherence Analysis

Coherence analysis is a method used to evaluate how meaningful and interpretable the topics generated by an LDA model are. It assesses the semantic similarity between the top words in each topic — topics with higher internal similarity (i.e., words that make sense together) score higher in coherence.

Notebook:

https://colab.research.google.com/drive/1ZKrtNUCwNbARTrVyE2xATsDZoFMgYdAm#scrollTo=8adf1b73-558b-49fb-9539-2bbbe9bd8278

3.3.1 Why Coherence Analysis?

- **Topic Quality Check:** Not all sets of topics produced by LDA are useful. Coherence helps us **filter out noisy or mixed topics**.

- **Determine Optimal Number of Topics:** By comparing coherence scores across multiple values of *k* (number of topics), we can identify which topic count yields the most interpretable results.

- **Supports Better Downstream Analysis:** Accurate and coherent topics improve the reliability of sentiment analysis grouped by topic.

3.3.1.1 How We Did It:

- We trained multiple LDA models with different values of **k** (number of topics), e.g., `k = 5, 7, 10`.

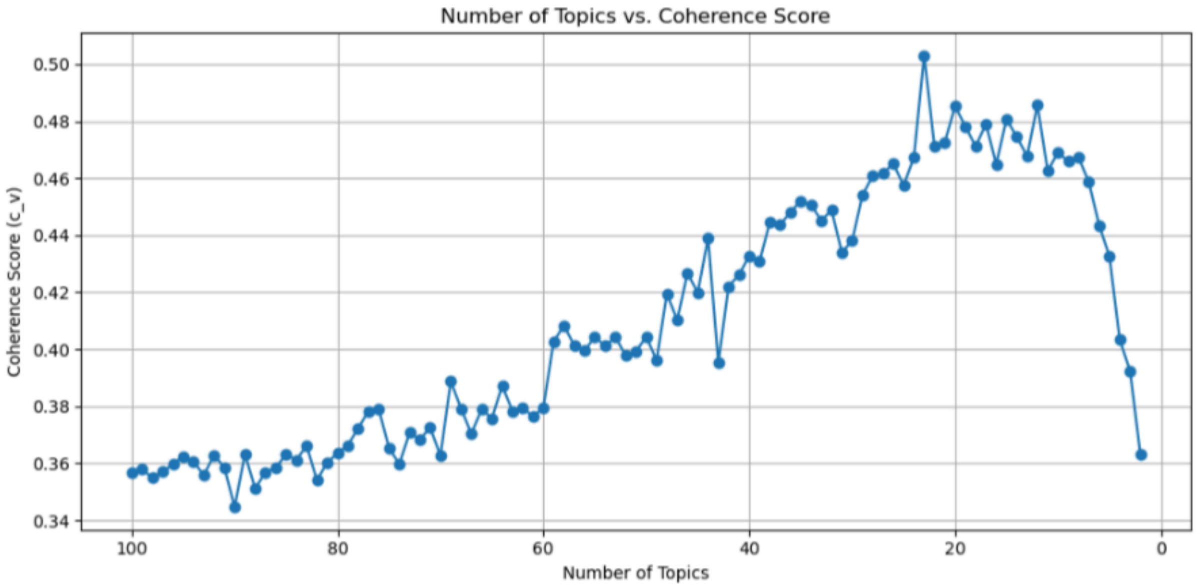
- For each model, we computed the **coherence score** using the `CoherenceModel` function from the **Gensim** library.

- Coherence type used: `c_v` (a popular choice for human interpretability).

- The model with the **highest coherence score** was selected as our final topic model.
- Stored and plotted the coherence scores to identify the topic count with the highest score.

```
coherence_model = CoherenceModel(

            model=lda,

            texts=tokenized_docs,

            dictionary=dictionary,

            coherence='c_v'

        )



        score = coherence_model.get_coherence()
```

3.3.2 GRAPH OF THE COHERENCE ANALYSIS



3.4 Probability Distributions

The LDA model assigns probabilities to each topic per review, helping identify dominant themes and their frequency.

Notebook:

https://colab.research.google.com/drive/1htz2AdWLa_KtiYOMknprpSJap9gH418-#scrollTo=PKa_XuyW4aMt

3.4.1 Dominant Topic

Each review was assigned:

- Its **dominant topic** based on the topic with the **highest probability** from the LDA topic distribution.

The LDA model gives a probability distribution over all topics for each review. The topic with the highest probability is considered the **dominant topic**, as it best represents the main theme of the review.

This step was essential for grouping reviews based on their main subject and performing topic-wise sentiment analysis.

3.4.2 Calculating Sentiment Distribution by Topic

3.4.2.1 Definition:

For each dominant topic, we calculated the **percentage of reviews** within that topic that were:

- **Positive**
- **Neutral**
- **Negative**

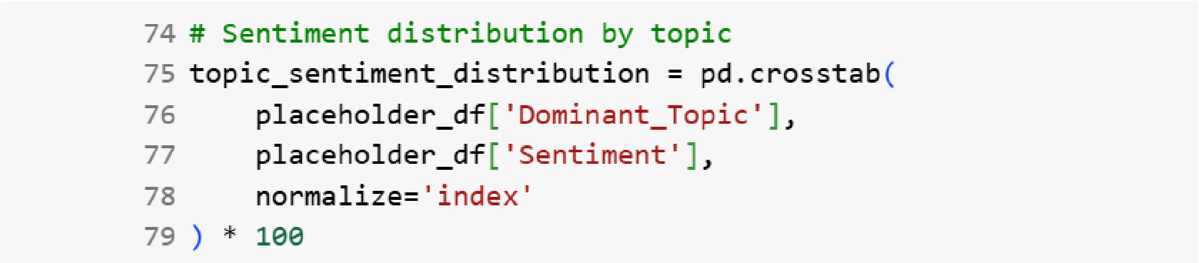
This gave us a **topic-wise sentiment profile**, helping us understand how customers feel about each major theme discussed.

3.4.2.2 Why?

- **Identify Drivers of Satisfaction:**
  Understand which topics (e.g., delivery, quality, support) are most associated with positive or negative experiences.
- **Targeted Improvements:**
  Helps businesses focus on topics with high **negative sentiment** and take corrective action.

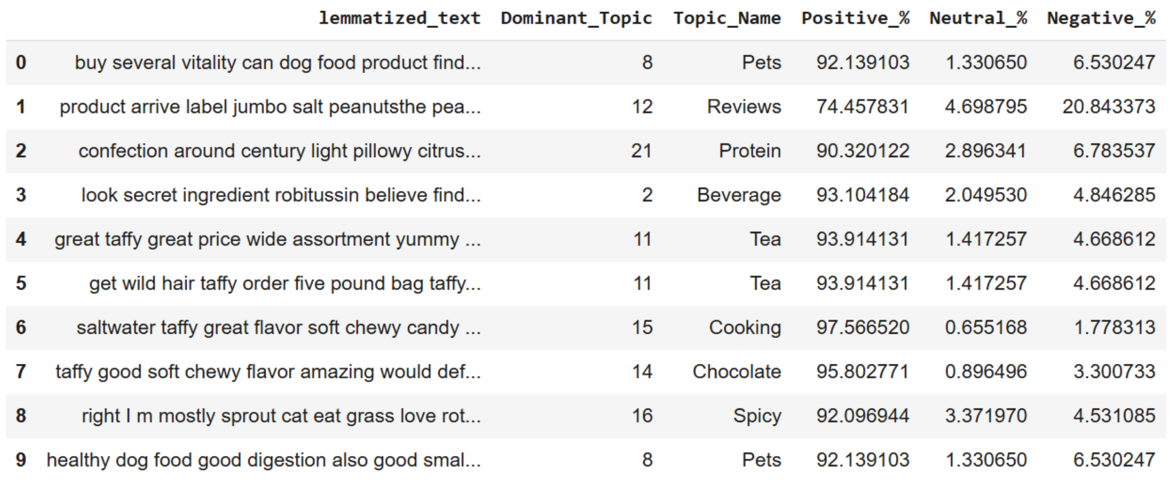
3.4.2.3 How We Did It:

1. **Grouped reviews** based on their dominant topic.
2. **Ran sentiment analysis** on each review using [VADER].
3. Assigned each review a **sentiment label** (positive, neutral, or negative).
4. Used `pandas.crosstab()` to calculate the **normalized distribution** of sentiments per topic.
   This gave us the **percentage of each sentiment class** per topic.

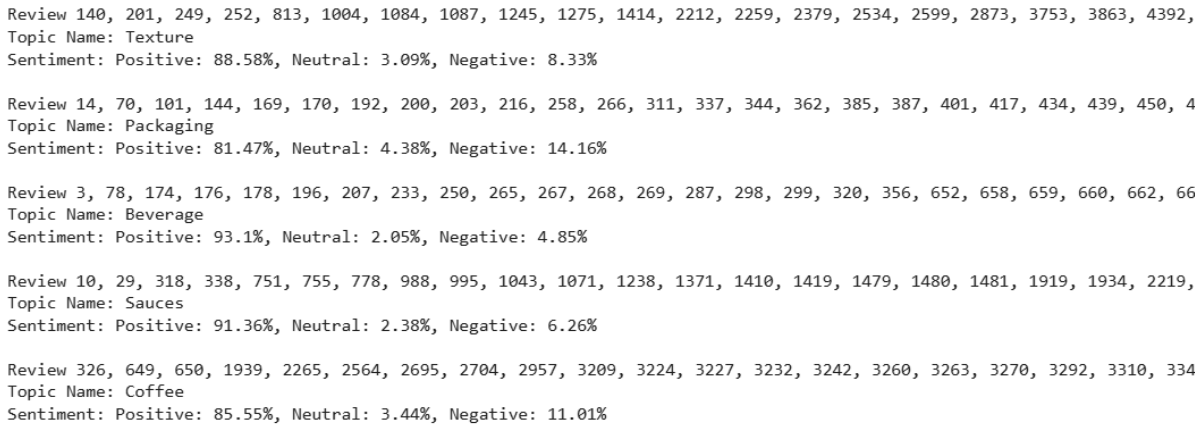


```
# Sentiment distribution by topic
topic_sentiment_distribution = pd.crosstab(
    placeholder_df['Dominant_Topic'],
    placeholder_df['Sentiment'],
    normalize='index'
) * 100
```

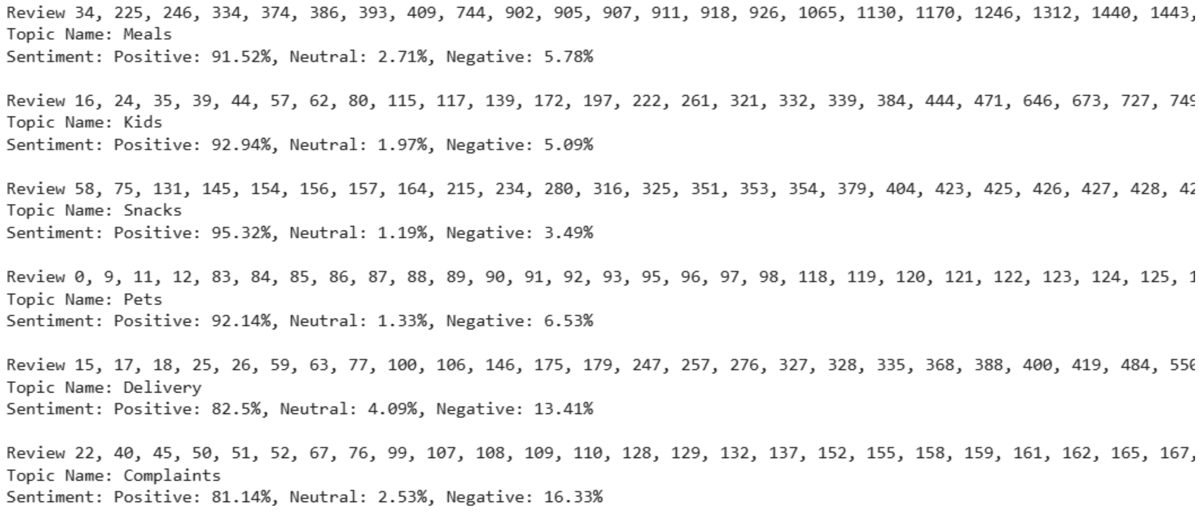
OUTPUT



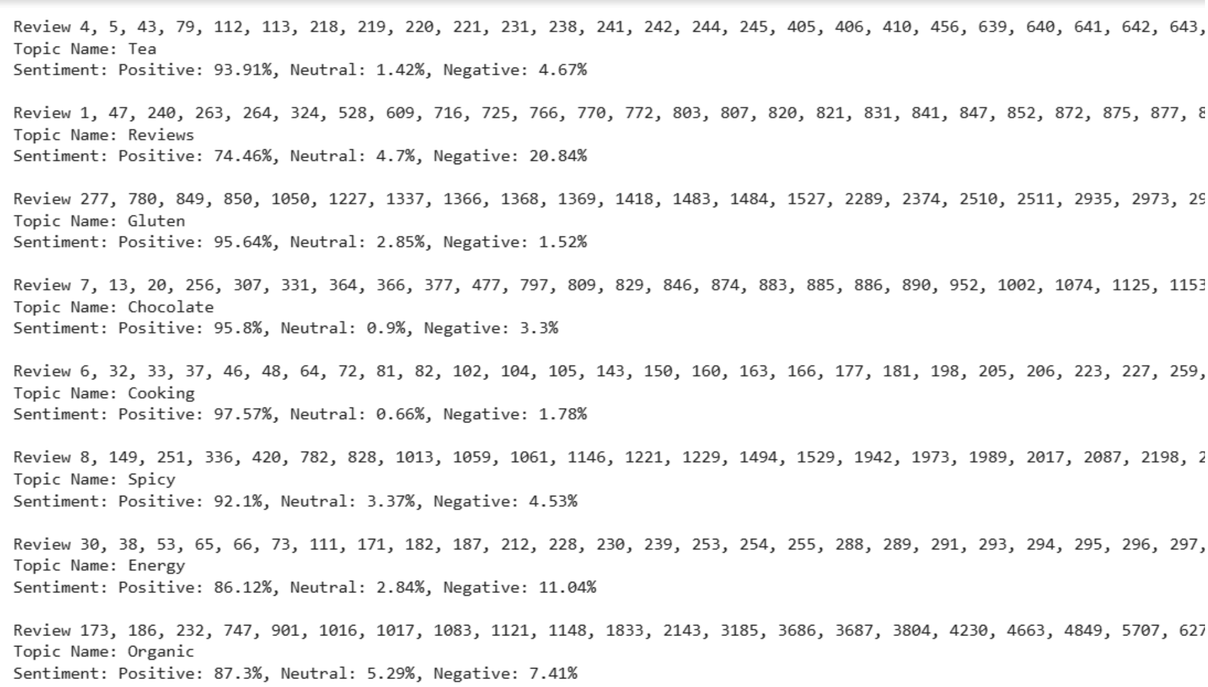
	lemmatized_text	Dominant_Topic	Topic_Name	Positive_%	Neutral_%	Negative_%
0	buy several vitality can dog food product find...	8	Pets	92.139103	1.330650	6.530247
1	product arrive label jumbo salt peanutsthe pea...	12	Reviews	74.457831	4.698795	20.843373
2	confection around century light pillowy citrus...	21	Protein	90.320122	2.896341	6.783537
3	look secret ingredient robitussin believe find...	2	Beverage	93.104184	2.049530	4.846285
4	great taffy great price wide assortment yummy ...	11	Tea	93.914131	1.417257	4.668612
5	get wild hair taffy order five pound bag taffy...	11	Tea	93.914131	1.417257	4.668612
6	saltwater taffy great flavor soft chewy candy ...	15	Cooking	97.566520	0.655168	1.778313
7	taffy good soft chewy flavor amazing would def...	14	Chocolate	95.802771	0.896496	3.300733
8	right I m mostly sprout cat eat grass love rot...	16	Spicy	92.096944	3.371970	4.531085
9	healthy dog food good digestion also good smal...	8	Pets	92.139103	1.330650	6.530247



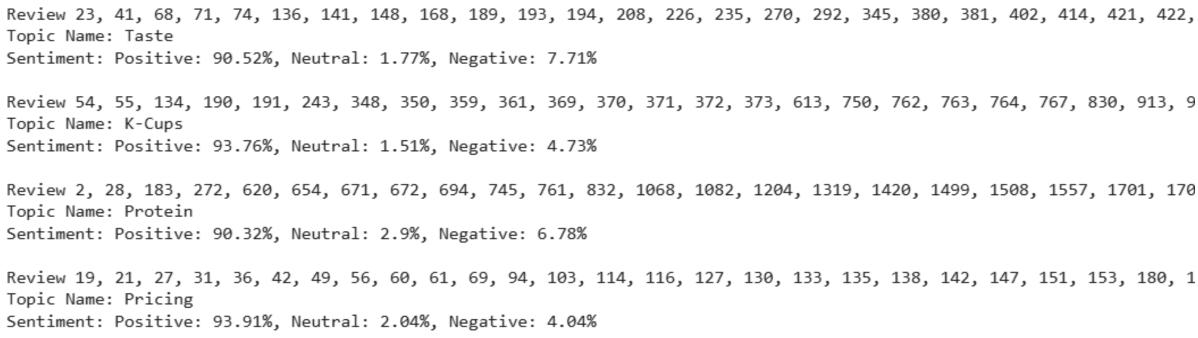
```
Review 140, 201, 249, 252, 813, 1004, 1084, 1087, 1245, 1275, 1414, 2212, 2259, 2379, 2534, 2599, 2873, 3753, 3863, 4392,
Topic Name: Texture
Sentiment: Positive: 88.58%, Neutral: 3.09%, Negative: 8.33%

Review 14, 70, 101, 144, 169, 170, 192, 200, 203, 216, 258, 266, 311, 337, 344, 362, 385, 387, 401, 417, 434, 439, 450, 4
Topic Name: Packaging
Sentiment: Positive: 81.47%, Neutral: 4.38%, Negative: 14.16%

Review 3, 78, 174, 176, 178, 196, 207, 233, 250, 265, 267, 268, 269, 287, 298, 299, 320, 356, 652, 658, 659, 660, 662, 66
Topic Name: Beverage
Sentiment: Positive: 93.1%, Neutral: 2.05%, Negative: 4.85%

Review 10, 29, 318, 338, 751, 755, 778, 988, 995, 1043, 1071, 1238, 1371, 1410, 1419, 1479, 1480, 1481, 1919, 1934, 2219,
Topic Name: Sauces
Sentiment: Positive: 91.36%, Neutral: 2.38%, Negative: 6.26%

Review 326, 649, 650, 1939, 2265, 2564, 2695, 2704, 2957, 3209, 3224, 3227, 3232, 3242, 3260, 3263, 3270, 3292, 3310, 334
Topic Name: Coffee
Sentiment: Positive: 85.55%, Neutral: 3.44%, Negative: 11.01%
```



```
Review 34, 225, 246, 334, 374, 386, 393, 409, 744, 902, 905, 907, 911, 918, 926, 1065, 1130, 1170, 1246, 1312, 1440, 1443,
Topic Name: Meals
Sentiment: Positive: 91.52%, Neutral: 2.71%, Negative: 5.78%

Review 16, 24, 35, 39, 44, 57, 62, 80, 115, 117, 139, 172, 197, 222, 261, 321, 332, 339, 384, 444, 471, 646, 673, 727, 749
Topic Name: Kids
Sentiment: Positive: 92.94%, Neutral: 1.97%, Negative: 5.09%

Review 58, 75, 131, 145, 154, 156, 157, 164, 215, 234, 280, 316, 325, 351, 353, 354, 379, 404, 423, 425, 426, 427, 428, 42
Topic Name: Snacks
Sentiment: Positive: 95.32%, Neutral: 1.19%, Negative: 3.49%

Review 0, 9, 11, 12, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 95, 96, 97, 98, 118, 119, 120, 121, 122, 123, 124, 125, 1
Topic Name: Pets
Sentiment: Positive: 92.14%, Neutral: 1.33%, Negative: 6.53%

Review 15, 17, 18, 25, 26, 59, 63, 77, 100, 106, 146, 175, 179, 247, 257, 276, 327, 328, 335, 368, 388, 400, 419, 484, 556
Topic Name: Delivery
Sentiment: Positive: 82.5%, Neutral: 4.09%, Negative: 13.41%

Review 22, 40, 45, 50, 51, 52, 67, 76, 99, 107, 108, 109, 110, 128, 129, 132, 137, 152, 155, 158, 159, 161, 162, 165, 167,
Topic Name: Complaints
Sentiment: Positive: 81.14%, Neutral: 2.53%, Negative: 16.33%
```



```
Review 4, 5, 43, 79, 112, 113, 218, 219, 220, 221, 231, 238, 241, 242, 244, 245, 405, 406, 410, 456, 639, 640, 641, 642, 643,
Topic Name: Tea
Sentiment: Positive: 93.91%, Neutral: 1.42%, Negative: 4.67%

Review 1, 47, 240, 263, 264, 324, 528, 609, 716, 725, 766, 770, 772, 803, 807, 820, 821, 831, 841, 847, 852, 872, 875, 877, 8
Topic Name: Reviews
Sentiment: Positive: 74.46%, Neutral: 4.7%, Negative: 20.84%

Review 277, 780, 849, 850, 1050, 1227, 1337, 1366, 1368, 1369, 1418, 1483, 1484, 1527, 2289, 2374, 2510, 2511, 2935, 2973, 29
Topic Name: Gluten
Sentiment: Positive: 95.64%, Neutral: 2.85%, Negative: 1.52%

Review 7, 13, 20, 256, 307, 331, 364, 366, 377, 477, 797, 809, 829, 846, 874, 883, 885, 886, 890, 952, 1002, 1074, 1125, 1153
Topic Name: Chocolate
Sentiment: Positive: 95.8%, Neutral: 0.9%, Negative: 3.3%

Review 6, 32, 33, 37, 46, 48, 64, 72, 81, 82, 102, 104, 105, 143, 150, 160, 163, 166, 177, 181, 198, 205, 206, 223, 227, 259,
Topic Name: Cooking
Sentiment: Positive: 97.57%, Neutral: 0.66%, Negative: 1.78%

Review 8, 149, 251, 336, 420, 782, 828, 1013, 1059, 1061, 1146, 1221, 1229, 1494, 1529, 1942, 1973, 1989, 2017, 2087, 2198, 2
Topic Name: Spicy
Sentiment: Positive: 92.1%, Neutral: 3.37%, Negative: 4.53%

Review 30, 38, 53, 65, 66, 73, 111, 171, 182, 187, 212, 228, 230, 239, 253, 254, 255, 288, 289, 291, 293, 294, 295, 296, 297,
Topic Name: Energy
Sentiment: Positive: 86.12%, Neutral: 2.84%, Negative: 11.04%

Review 173, 186, 232, 747, 901, 1016, 1017, 1083, 1121, 1148, 1833, 2143, 3185, 3686, 3687, 3804, 4230, 4663, 4849, 5707, 627
Topic Name: Organic
Sentiment: Positive: 87.3%, Neutral: 5.29%, Negative: 7.41%
```



```
Review 23, 41, 68, 71, 74, 136, 141, 148, 168, 189, 193, 194, 208, 226, 235, 270, 292, 345, 380, 381, 402, 414, 421, 422,
Topic Name: Taste
Sentiment: Positive: 90.52%, Neutral: 1.77%, Negative: 7.71%

Review 54, 55, 134, 190, 191, 243, 348, 350, 359, 361, 369, 370, 371, 372, 373, 613, 750, 762, 763, 764, 767, 830, 913, 9
Topic Name: K-Cups
Sentiment: Positive: 93.76%, Neutral: 1.51%, Negative: 4.73%

Review 2, 28, 183, 272, 620, 654, 671, 672, 694, 745, 761, 832, 1068, 1082, 1204, 1319, 1420, 1499, 1508, 1557, 1701, 170
Topic Name: Protein
Sentiment: Positive: 90.32%, Neutral: 2.9%, Negative: 6.78%

Review 19, 21, 27, 31, 36, 42, 49, 56, 60, 61, 69, 94, 103, 114, 116, 127, 130, 133, 135, 138, 142, 147, 151, 153, 180, 1
Topic Name: Pricing
Sentiment: Positive: 93.91%, Neutral: 2.04%, Negative: 4.04%
```

Topic_No	Topic_Name	Positive%	Neutral%	Negative%
0	Texture	88.58%	3.09%	8.33%
1	Packaging	81.47%	4.38%	14.16%
2	Beverage	93.1%	2.05%	4.85%
3	Sauces	91.36%	2.38%	6.26%
4	Coffee	85.55%	3.44%	11.01%
5	Meals	91.52%	2.71%	5.78%

6	Kids	92.94%	1.97%	5.09%
7	Snacks	95.32%	1.19%	3.49%
8	Pets	92.14	1.33%	%6.53%
9	Delivery	82.5%	4.09%	13.41%
10	Complaints	81.14%	2.53%	16.33%
11	Tea	93.91%	1.42%	4.67%
12	Reviews	74.46%	4.7%	20.84%
13	Gluten	95.64%	2.85%	1.52%
14	Chocolate	95.8%	0.9%	3.3%
15	Cooking	97.57%	0.66%	1.78%
16	Spicy	92.1%	3.37%	4.53%
17	Energy	86.12%	2.84%	11.04%
18	Organic	87.3%	5.29%	7.41%
19	Taste	90.52%	1.77%	7.71%
20	K-Cups	93.76%	1.51%	4.73%
21	Protein	90.32%	2.9%	6.78%
22	Pricing	93.91%	2.04%	4.04%

6. Observation

The results revealed how different topics align with customer sentiments and which areas impact satisfaction the most.

6.1 Highly Positive Topics

- **Cooking (97.6% positive):**
  Reviews in this topic are almost universally positive, indicating high satisfaction with cooking-related products (e.g., ingredients, utensils, recipes).
- **Chocolate (95.8% positive):**
  Chocolate products are well-loved, with minimal negative feedback.
- **Tea, Beverage, Protein, Spicy:**
  These topics also show very high positive sentiment, reflecting strong customer approval.

Implication:

Products in these categories are meeting or exceeding customer expectations. They can be leveraged for positive marketing and as models for other product lines.

3.4.3 Mixed or Negative-Leaning Topics

- **Reviews (20.8% negative):**
  This topic has a notably higher share of negative sentiment, indicating that reviews discussing product feedback, complaints, or comparisons often express dissatisfaction.
- **Pets (6.5% negative):**
  While still mostly positive, pet-related products have a slightly higher negative rate than, say, chocolate or tea, possibly due to the sensitive nature of pet health and preferences.

Implication:

Topics with higher negative sentiment highlight areas where customer expectations are not being met. “Reviews” may include critical feedback on product quality, shipping, or customer service

3.5 Patterns and Trends

Highly positive sentiments were observed around product quality, while issues like delivery and packaging showed more negative feedback.

3.5.1 Sentiment Concentration

- Positive sentiment is highly concentrated in topics related to taste, enjoyment, and staple foods.
- Negative sentiment is not evenly distributed; it clusters in topics where customers discuss broader experiences, such as reviews or complaints, rather than specific products.

3.5.2 Neutral Sentiment

Neutral sentiment is generally low across all topics, suggesting that customers tend to express clear opinions (either positive or negative) rather than ambivalence.

3.6 Conclusion

The topic-sentiment analysis provides a nuanced, actionable map of customer satisfaction across product themes. Most products delight customers, but specific areas (notably those involving critical reviews or complaints) reveal opportunities for improvement. This approach empowers businesses to move from anecdotal feedback to data-driven strategy, closing the loop between customer voice and business action.

7 Analysis of Factors Impacting Customer Satisfaction

The second part of this study focuses on identifying and interpreting the key factors that influence customer satisfaction, as expressed in their reviews.

7.1 Definition of Factors

Factors refer to the main topics or themes found within customer reviews. These represent different aspects of the product or service experience, such as:

- Product taste
- Packaging quality
- Delivery experience
- Pricing
- Customer service
- Specific product categories (e.g., tea, snacks)

These factors were extracted through topic modeling of the review text.

3.6.0.1 Identification of Factors

4 To discover these factors, Latent Dirichlet Allocation (LDA) was applied to the customer reviews dataset. LDA automatically identifies dominant topics in the text by grouping commonly co-occurring words. Each topic was then manually assigned a clear, descriptive name—like “Packaging,” “Complaints,” or “Chocolate”—to facilitate interpretation.

4.0.0.1 Measuring Customer Satisfaction

5 Customer satisfaction was measured by analyzing the sentiment of each review using VADER sentiment analysis. Reviews were categorized as Positive, Neutral, or Negative based on their sentiment scores. For each identified factor, the percentage of reviews falling into each sentiment category was calculated. For example:

- Reviews about **Delivery** showed approximately 82.5% positive sentiment

- Reviews about **Cooking** showed 97.6% positive sentiment

8 This approach quantifies customer feelings towards each factor.

8.0.0.1 Analysis of Findings

9 The sentiment distributions reveal which factors contribute to customer satisfaction or dissatisfaction:

- Factors like **Cooking**, **Chocolate**, and **Tea** had predominantly positive sentiment, indicating strong customer satisfaction.

- Factors such as **Complaints**, **Reviews**, and **Delivery** showed higher negative sentiment, highlighting areas causing dissatisfaction.

12 These patterns help pinpoint what aspects customers value most and where improvements are needed.

12.0.0.1 Business Implications

13 These insights offer actionable guidance for businesses:

- Focus on improving weaker areas such as delivery and packaging to reduce dissatisfaction.

- Leverage strengths like taste and product quality to boost customer loyalty.

- Make data-driven decisions based on actual customer feedback rather than assumptions.

17 By targeting efforts effectively, companies can enhance overall customer satisfaction and improve business outcomes.

18 Conclusion and Future Work

The project highlights valuable insights for business improvement, with scope for advanced modeling, real-time dashboards, and multilingual support in the future.

18.1 Summary of Key Findings

Through rigorous data preprocessing, advanced topic modeling, and robust sentiment analysis, our project has yielded several important insights into customer satisfaction within the Amazon Fine Food Reviews dataset:

- **Dominance of Positive Sentiment:**
  The vast majority of reviews across most product topics express positive sentiment, indicating overall customer satisfaction with the products analyzed.

- **Concentration of Negative Sentiment:**
  Negative sentiment is not uniformly distributed. It is concentrated in specific topics, particularly those related to general feedback, complaints, and product comparisons (“Reviews” topic).

- **Topic-Specific Satisfaction:**
  Certain product categories, such as “Cooking,” “Chocolate,” “Tea,” and “Beverage,” exhibit exceptionally high positive sentiment, suggesting these products consistently meet or exceed customer expectations.

- **Low Neutral Sentiment:**

Most customers express clear opinions, with neutral sentiment forming a small proportion of reviews. This suggests that customers are motivated to leave feedback primarily when they have strong positive or negative experiences.

- **Actionable Patterns:**
  The joint analysis of topics and sentiments provides a roadmap for identifying both strengths to leverage and weaknesses to address within the product portfolio.

18.2 Potential Future Work and Improvements

While our analysis provides a solid foundation, several avenues exist for further enhancing the analytical framework:

1. **Incorporate Advanced NLP Models:**
   a. Fine-tune transformer-based models (e.g., BERT, RoBERTa) on domain-specific review data for improved sentiment accuracy, especially in detecting sarcasm, irony, and nuanced opinions.
2. **Expand Topic Modeling:**
   a. Apply hierarchical topic modeling or BERTopic to capture sub-topics and more granular themes within broad topics.
   b. Explore dynamic topic modeling to track how topics and sentiments evolve over time.
3. **Address Data Imbalance:**
   a. Experiment with data augmentation or re-sampling techniques to mitigate the skew towards positive reviews and enhance the detection of minority sentiments.
4. **Integrate Multimodal Data:**
   a. Combine textual reviews with other data sources (e.g., star ratings, product metadata, images) for a more holistic analysis of customer satisfaction.
5. **Real-Time and Multilingual Analysis:**
   a. Develop real-time sentiment dashboards for business users.
   b. Extend the framework to analyze reviews in multiple languages, broadening its applicability to global markets.
6. **User Segmentation and Personalization:**
   a. Analyze sentiment and topic trends by customer segments (e.g., location, purchase frequency) to enable targeted marketing and service.

18.3 Final Thoughts

This project demonstrates the power of combining modern NLP techniques with statistical analysis and visualization to transform raw customer feedback into actionable business intelligence. By systematically uncovering what customers care about and how they feel, organizations can make informed decisions that enhance product offerings, improve customer experience, and drive business growth.

The analytical pipeline established here is scalable and adaptable, offering a blueprint for sentiment and topic analysis in a wide range of domains beyond food reviews. As customer feedback continues to grow in volume and importance, such data-driven approaches will be essential for maintaining competitive advantage and fostering long-term customer loyalty.