(Automated Customer Reviews)

1st ESSA Al-HASSAN  
 2st *RAZAN BAHHARI*   
3st *GHALA AL-OTIBIE*



*Abstract*— As part of the "Automated Customer Reviews" process, we are developing an integrated system that uses artificial intelligence techniques to analyze and evaluate customer reviews. The main goal is to classify reviews into positive, negative, and neutral categories based on star ratings, using pre-trained models such as BERT and RoberTa. Product categories are then grouped into 4-6 main categories, such as "Electronics" and "Accessories," based on in-depth data analysis.

# **Introduction**

As part of the "Automated Customer Reviews" project, we trained advanced AI-based models to read and analyze customer reviews. We began by categorizing reviews into positive, negative, and neutral categories, using pre-trained models like BERT and RoberTa, which helped us understand customer sentiment more accurately.

After categorizing the reviews, we grouped product categories into 4-6 main categories, such as "Electronics" and "Accessories," based on a comprehensive data analysis. This grouping helped simplify the information and make it easier to access.

In addition, we used models like T5 and GPT-3 to summarize the reviews, allowing us to create short articles recommending the best products in each category, highlighting key complaints and less-favorite products.

Finally, we created an interactive website that enables users to easily explore reviews and recommendations, enhancing the shopping experience and increasing trust in products. This process is an important step toward improving companies' understanding of their customers' needs and promoting transparency in the market.

***A. PROBLEM STATEMENT.***

In today's digital marketplace, customer reviews play a crucial role in influencing purchasing decisions. However, the sheer volume of these reviews makes it challenging for businesses to extract actionable insights. Many companies struggle to effectively classify, aggregate, and summarize customer feedback, leading to missed opportunities for improvement and customer satisfaction.

Traditional methods of analyzing reviews are often time-consuming and insufficient for understanding nuanced sentiments. This gap in effective data analysis hampers the ability of businesses to respond quickly to customer needs and adapt their products or services accordingly.

***B. PROPOSED SOLUATION.***

We propose an automated system that uses advanced Natural Language Processing (NLP) techniques to efficiently analyze customer reviews.

***C. Project goal.***

This project aims to develop a product review system powered by NLP models that aggregate customer feedback from different sources. The key tasks include classifying reviews, clustering product categories, and using generative AI to summarize reviews into recommendation articles.

# ***DATA PROCESSING***

We used three datasets (1429\_1 ,

datafiniti\_Amazon\_consumer\_Reviews\_of\_Amazon\_

product , datafiniti\_Amazon\_consumer\_Reviews\_of\_

Amazon\_product ) merged them, and named each dataset (new\_data) because it was the latest version of the dataset. We used it for classification, aggregation, and summarization. We took this step because a single dataset was not sufficient for some of the steps mentioned above. After that, it was read, cleaned, and the requirements were applied as required. We took this step for several reasons:

## **Reasons for choosing to combine** : Several datasets were combined to form a large-scale dataset so that we could use it in all the models that were applied and also so that we could use new and existing data and extract the data we wanted.

## **Challenges I Faced:** During the data merging phase, we had two important steps in this project: data partitioning and data balancing. We had to implement these two steps, but we were supposed to split the data into trained data and validation data before balancing them. This challenge we faced, and we had to merge the data to have a larger set. If we split it into training and validation data, the amount of data would be smaller, making it difficult to balance the data. We also had an additional point: dividing the data into negative, positive, and neutral data. This was the problem, as balancing them was difficult because they relied solely on the training portion of the data.

## **The most important columns in this project:**

## The columns that were important in the classification task were the review texts and the review rate.

## In the cluster task, the three important columns were the name, the category, and the review text. We used them based on the results that were likely to be highly accurate.

# **TASKS**

**1- CLASSIFICATION**

**Approach:** The classification approach in machine learning involves assigning predefined labels to input data based on its characteristics. Here is a structured diagram of this approach:

**A- Problem Definition:**

Clearly define the classification task, including the type of input data and the labels to be predicted. For example, we have classified customer reviews as positive, negative, or neutral.

**B- Data Collection:**

Collect a relevant dataset containing the input attributes and their corresponding labels. The dataset must be large enough to train a robust model.

**C- Data Preprocessing :**

Clean the data by addressing missing values, removing duplicates, normalizing or standardizing attributes, and selecting appropriate columns for classification (e.g., review text and review rate).

**C- Data Splitting:**

Split the data into two parts (one for training data and the other for validation).

For text data, we applied the TOKNAIZER technique after split data to (review and rate ) now we use split it using a balance method because In order to equate positive, negative and neutral data .

**Model:** Roberta-base: It is dedicated to the BERT (Bidirectional Encoder Representations from Transformers) model.

**2- CLUSTRING**

**Approach**: method used to cluster and analyze text, particularly in the context of customer reviews. Here are the steps we implemented:

**A - Data Collection:**

We collected a wide range of customer reviews from multiple sources to ensure data diversity.

We selected three columns that represent how customer reviews are categorized (review text, categories, and names). After collecting them, we combined them and added a new column to the dataset that included each one.

**B- Data Processing:**

Data cleaning by removing noise, such as unnecessary tokens or common stop words.

We converted the text into a format suitable for analysis and removed duplicate values ​​that were not needed .

**C- Data Representation:**

We used text representation techniques such as Word Embeddings (TSNE) to convert text into numerical vectors.

**D- Cluster Analysis:**

We applied clustering algorithms such as K-means to identify patterns and clusters within the data.

**Model** : **sentence-transformers/all-MiniLM-L6-v2** :

It is designed to transform sentences into numerical representations (embeddings) that can be used in a variety of natural language processing (NLP) tasks.

**3- SUMMRIZATION**

**Approach:** The goal of the summarization task is to create article-style summaries of customer reviews across different product categories. Each article includes:

• The top 3 products in a given category, along with their most notable customer reviews and complaints.

• A summary of the worst-rated product and why it should be avoided.

**A- Dataset Used:**

For this task, I used the balance\_reviews\_dataset, which was originally prepared and refined during the review classification task. This dataset includes:

• Product names

• Main categories and subcategories

• Star ratings

• Review text

• Pre-sorted sentiment categories (positive, negative, or neutral).

**B- Reviews Summarization:**

For each product, I summarized the positive and negative reviews separately.

The reviews were then grouped by sentiment and compiled into a single text input for the summarization model.

**C- Article Generation:**

The summaries were used to generate blog-style articles for each category.

The articles described the best and worst products, highlighting customer preferences and complaints.

**Model**: **t5-small** : It is an advanced Natural Language Processing (NLP) model based on the Transformers architecture.

# **DEPLOYMENT**

**1. Deployment Task Overview:**

This project is a web-based application built using Gradio and Hugging Face Transformers to analyze Amazon product reviews. The goal is to assist marketing teams and decision-makers by automatically:

Classifying customer sentiment (positive, neutral, or negative)

Clustering product types into categories

Generating article-style summaries for each category or product

The system helps users gain instant insights into how products are perceived by customers, what common issues exist, and which products perform best in each category.

**2. Datasets and Preprocessing:**

A- Datasets Used:

1- clustered\_reviews.csv: Contains Amazon reviews with precomputed clustering labels, product names, and categories.

2- balanced\_reviews\_dataset.csv:

A dataset that balances review sentiment classes to support effective sentiment classification.

B- Preprocessing Steps:

1-Loaded both datasets and checked for missing values.

2-Renamed inconsistent columns for merging (e.g., reviews.text to text, product\_name to name).

3-Created a unified dataset with consistent column names: name, text, categories, cluster\_name.

4-Cleaned the text fields by removing extra commas, line breaks, and duplicate reviews.

5-Limited the number of reviews processed per product to speed up model inference.

**3. Task Approach:**

A. Sentiment Classification :

Model Used: roberta\_model (fine-tuned for text classification)

Purpose: Classifies each review as Positive, Neutral, or Negative.

Implementation:

Tokenize review text using AutoTokenizer

Pass input to AutoModelForSequenceClassification

Use model logits to determine predicted sentiment label

B. Product Category Clustering

Precomputed Offline: Using KMeans on sentence embeddings

Cluster Labels: Mapped to descriptive categories (e.g., Ebook Readers, Batteries, Accessories)

Usage: Enables users to search by category or view group-level insights

C. Review Summarization

Model Used: t5\_small (pretrained summarization model)

Purpose: Generates brief, coherent summaries of product feedback

Steps:

Clean and deduplicate review text

Limit to 3 representative reviews per product

Format a prompt: "Summarize reviews for 'Product X'..."

Use Hugging Face pipeline to generate summary

Repeat for top 3 products in category and one worst-rated product

**4. Web Application (User Interface):**

Framework: Gradio (Blocks)

Theme: Soft layout with Amazon branding

Layout Features:

Amazon logo, large headline, and call-to-action

Input field for product/category search

Accordion-style result display for clean user experience

Output boxes for matched product, category, sentiment stats, and AI summary

Styling:

Integrated TailwindCSS-like appearance

Used Gradio markdown and column components for layout

Icons and color highlights enhance readability and engagement

**5. Challenges Faced:**

Challenge and Solution :

Ch1- CSV file inconsistencies (column names, formatting)

Sol1- Added column renaming and missing field handling

Ch2- Summarization producing messy or repetitive output

Sol2- Cleaned inputs, reduced to 3 unique reviews, formatted prompts

Ch3- UI felt too basic for end-users

Sol3- Applied Tailwind styling and branding through custom Markdown

Ch4- Model performance (slow on large inputs)

Sol4- Limited reviews to 3–5, optimized with cleaner prompts

**6. Final Features Summary:**

🔍 Search by product or category name

🧾 Return the matched product and cluster category

📊 Visualize the sentiment distribution from classified reviews

📝 Generate a summary with:

Top 3 products in category (with complaint count)

AI summary for each

Worst product with generated summary

# **figur**

**1- CLASSIFICATION**

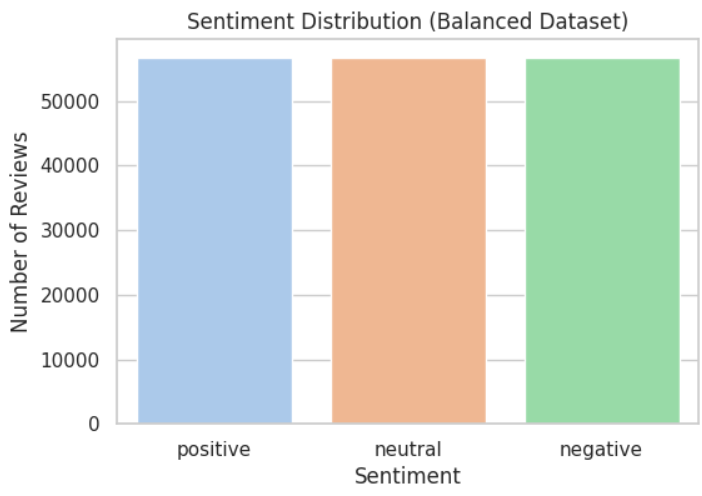


Fig. 1.1: balance

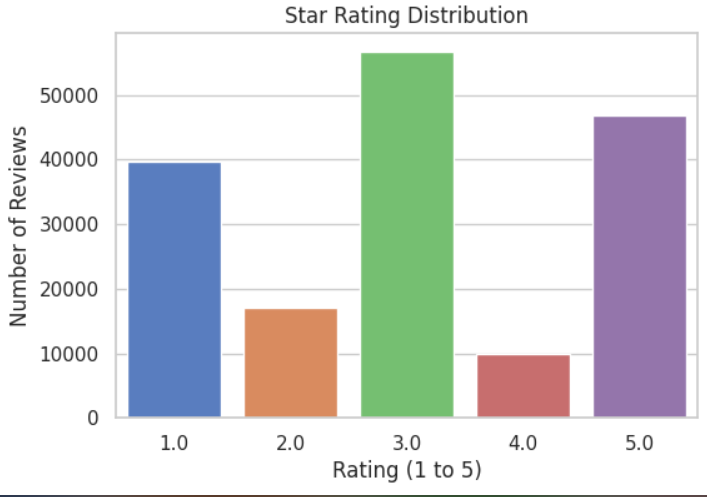


Fig. 1.2: Star Ratings



Fig. 1.3: Word Cloud : Most common words in a set of reviews

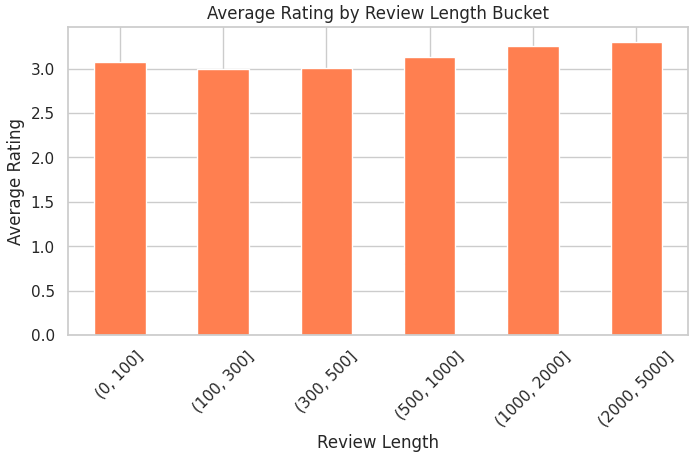


Fig. 1.4: Average rating by review length categories

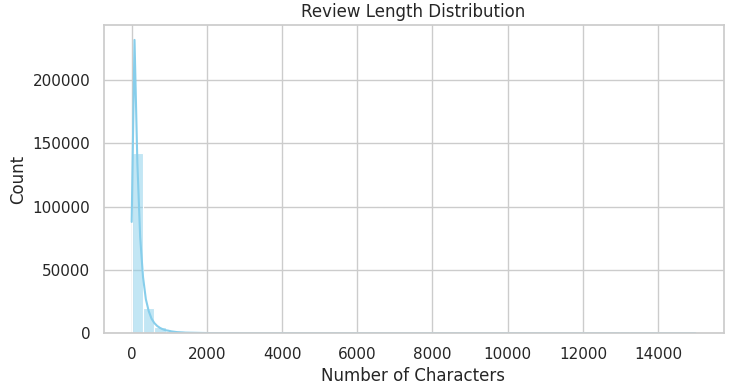


Fig. 1.5: Distribution of review length by number of characters

**2- CLUSTRING**

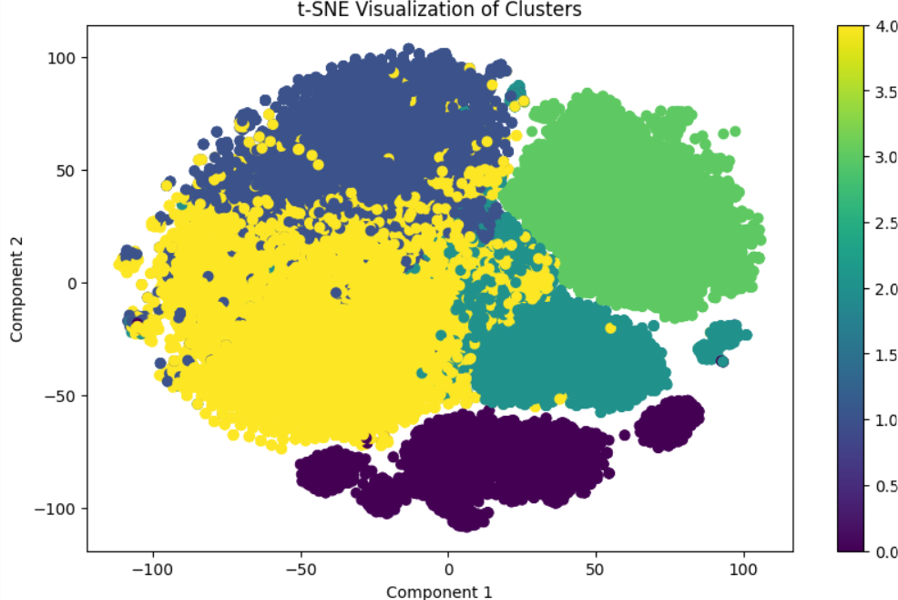


Fig. 2.1: Embedding(TSNE)

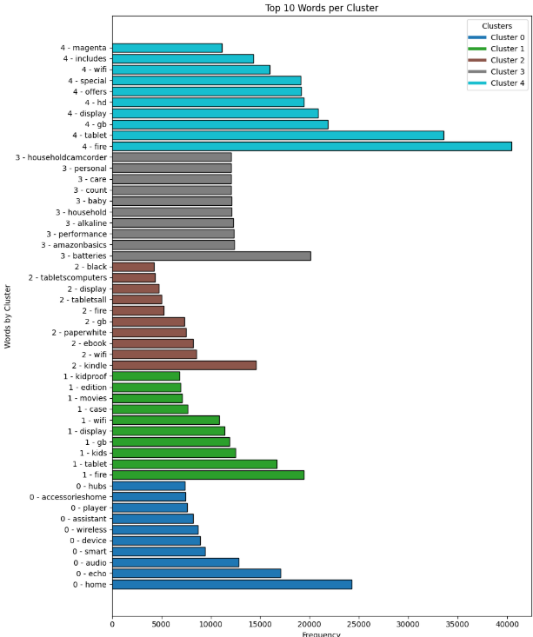


Fig. 2.2: TOP 10 WORDS (A collection of the most common words for each gathering)

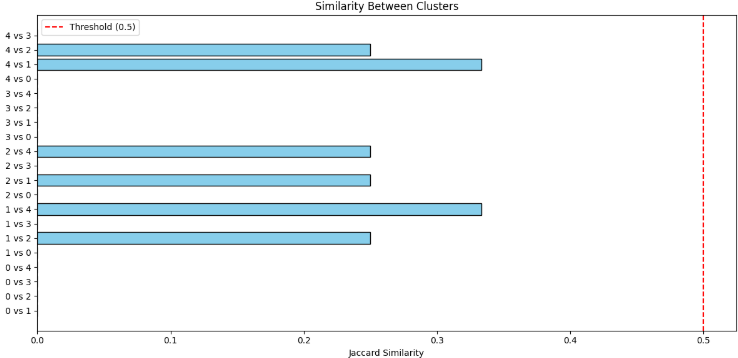


Fig. 2.3: Representing similarity between clusters using the Jaccard Similarity measure.

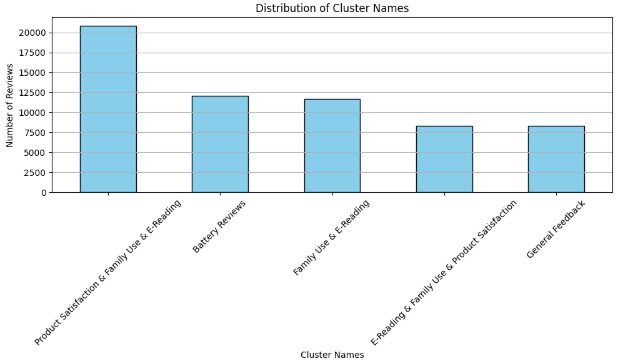


Fig. 2.4: Distribution of the number of reviews by cluster names after merging

**3- DEPLOYMENT**

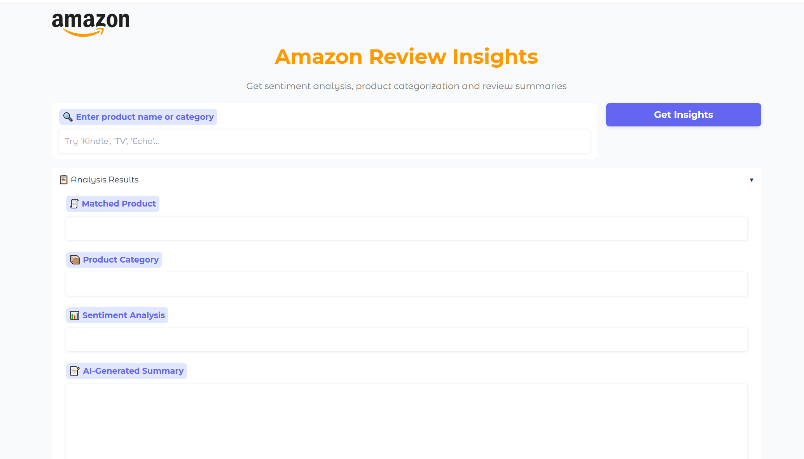


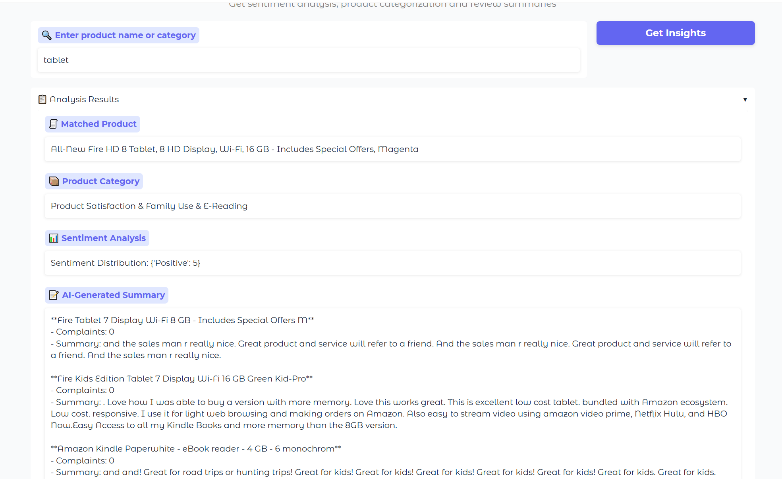
Fig. 3.1: Explanation of the interface of the tool for analyzing Amazon product reviews, with some of the key components 

Fig. 3.2: Product review analysis results using a data analysis tool

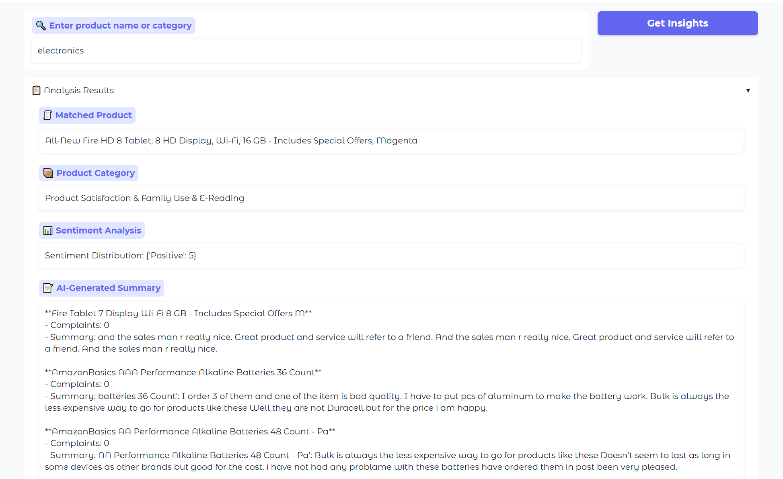


Fig. 3.3: Product reviews analysis results in the "Electronics" category

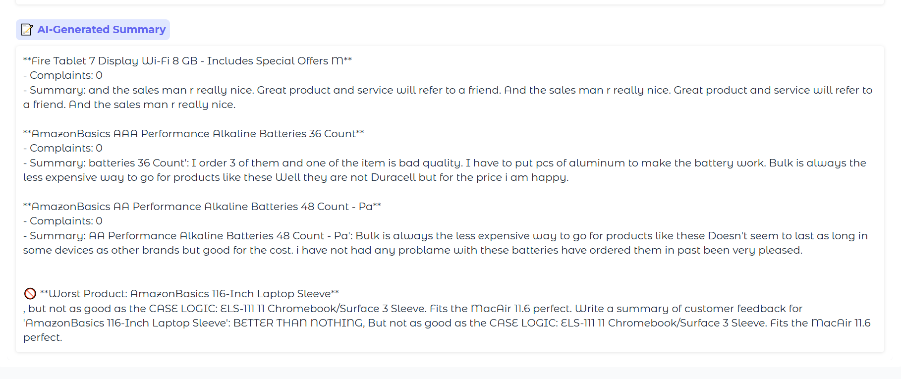


Fig. 3.4: An AI-generated summary of product reviews in the Electronics category.

# VI **CONCLUSION**

This project demonstrates a complete end-to-end NLP solution that combines sentiment analysis, product clustering, and generative AI to produce business insights from raw customer reviews. It not only showcases the power of pretrained transformer models but also emphasizes the importance of UI/UX design in making data science tools usable and impactful.

VII. **Future**

Research will continue to develop larger models like GPT-4 and T5, with improvements in performance and efficiency. These models will become more capable of understanding complex contexts.

**IX. References**

NLP Techniques:

NLP enables the analysis of large volumes of customer reviews by extracting sentiments, identifying trends, and categorizing feedback. Techniques such as sentiment analysis, topic modeling, and named entity recognition are commonly employed to derive meaningful insights from textual data [1].