PRODUCT RECOMMENDATION

AND

SENTIMENT ANALYSIS SYSTEM

GROUP 5

https://github.com/joakimTI/JUMIA-PRODUCTS-RECOMMENDATION-AND-PRODUCT-REVIEW-SYSTEM

OVERVIEW

1 Introduction

2 Problem Statement

3 Business Understanding

4 Data Understanding

5 Data Preparation

6 Data Analysis

7 Modelling

8 Results

9 Challenges

10 Conclusion

INTRODUCTION

E-commerce in Kenya has experienced remarkable growth, with platforms like Jumia drawing a significant user base. However, navigating vast and often poorly organized product catalogs poses challenges for users. Additionally, the lack of robust recommendation systems and unreliable reviews diminishes the shopping experience, leading to decreased user satisfaction

To address these gaps, we propose a recommendation and sentiment analysis system that harnesses user interaction data and feedback. This solution aims to deliver tailored product suggestions and actionable customer insights, enhancing user satisfaction.

PROBLEM STATEMENT

Inadequate product recommendations based on user behavior

This results in a lack of personalization, which diminishes the overall shopping experience.
This lack of personalization can lead to frustration, decreased engagement, and missed sales opportunities. For businesses, it translates into lower conversion rates, reduced customer retention, and less effective marketing strategies.

Unreliable rating and review system

This creates significant challenges for customers seeking to make informed purchasing decisions. Without dependable reviews, customers are left with limited or misleading insights into product quality, usability, and value.

Limited retailer insights into customer sentiment

This prevents retailers from fully understanding how their products and services resonate with users. This lack of awareness makes it difficult to identify areas for improvement, refine product offerings, or tailor marketing strategies to align with customer preferences..

BUSINESS UNDERSTANDING

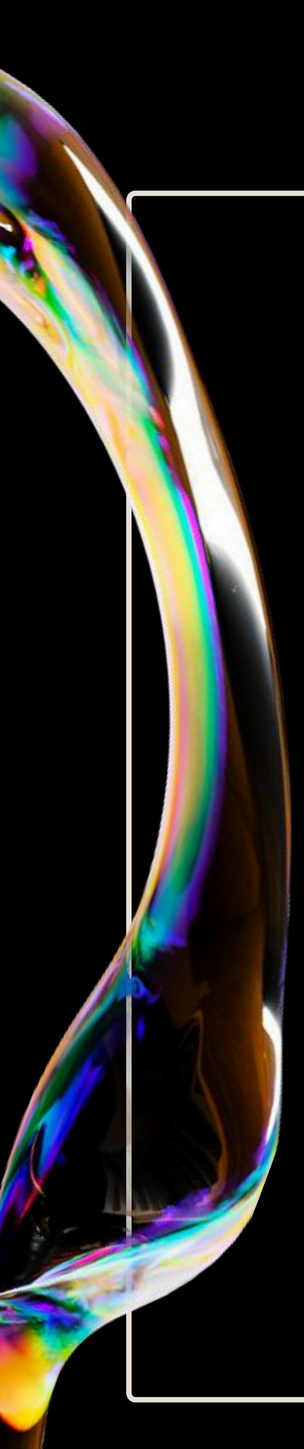
Our target audience are:

- **Primary Users (Consumers):** Customers who frequently shop on Jumia and need relevant recommendations to streamline their product search and improve purchase decisions.
- **Retailers/Sellers:** Businesses and individual sellers on Jumia who seek insights into customer preferences and sentiment to better tailor their product offerings and marketing strategies.
- **Jumia Management:** The platform administrators who aim to improve customer satisfaction, engagement, and conversion rates through improved site functionality.

DATA UNDERSTANDING

For this project, we used a dataset we scrapped from **Jumia**. The dataset includes product reviews and ratings, which are essential for building both the sentiment analysis and recommendation system models.

The relevant features, such as ‘user name’, ‘product name’, ‘category’, ‘review’ and ‘ratings’, are well-defined to facilitate both sentiment classification and recommendation tasks.

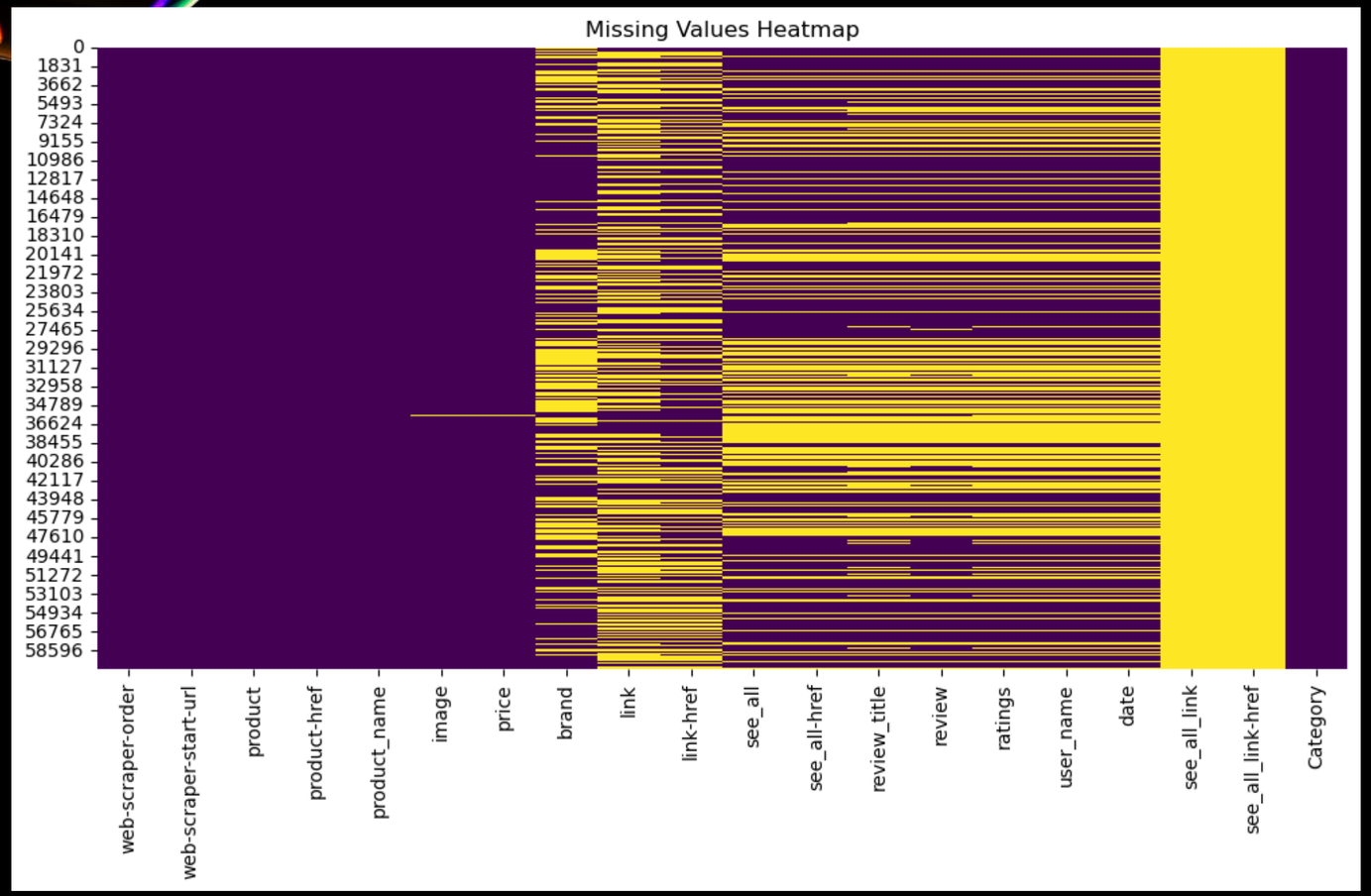


DATA PREPARATION

To prepare the data for modelling, these are the steps we took:

- **Handling missing values:** Identify and address gaps in the data by imputing missing values and removing incomplete records to ensure consistency
- **Feature selection and engineering:** Select the most relevant variables and created new ones to improve model performance and highlight underlying patterns.
- **One-hot encoding and standardization:** Convert categorical variables into binary columns and scale numerical features to a uniform range for better compatibility with our machine learning algorithms.

DATA CLEANING



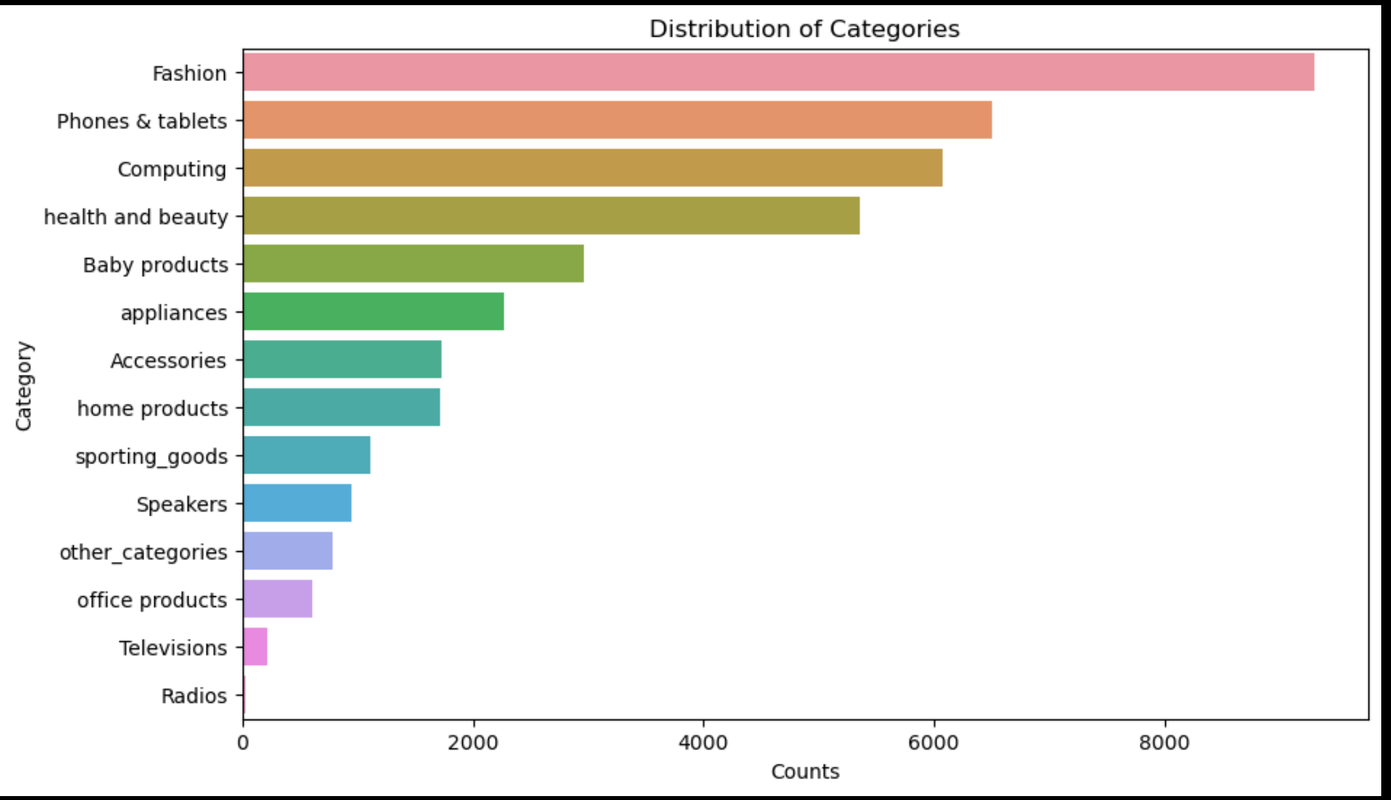
MISSING VALUES

This heatmap provides insight to the following:

- Which columns need data cleaning
- Columns that might need to be dropped or imputed
- Potential systematic issues in data collection

DATA ANALYSIS

CATEGORIES



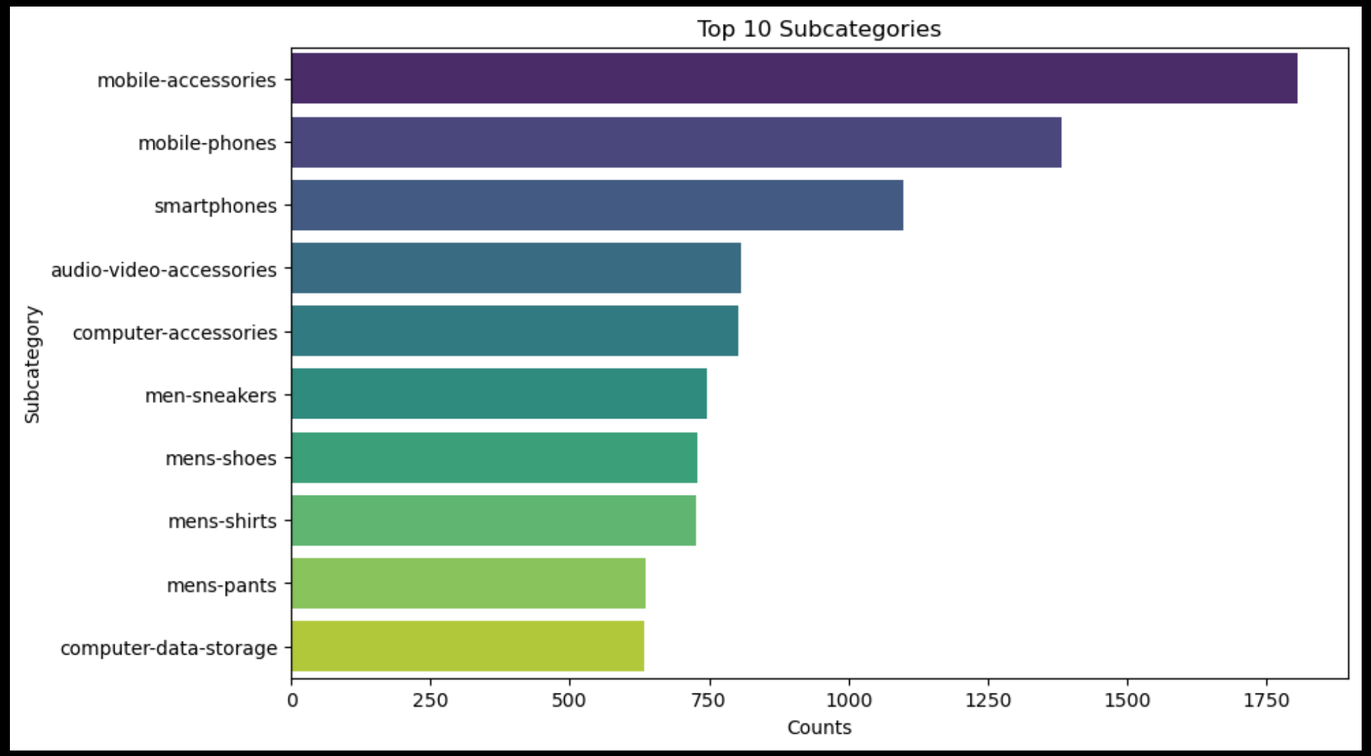
The distribution of categories provide the following insights:

Customer Interest: The top categories reflect higher customer interest or demand

Inventory Planning: Retailers could focus more on stocking and promoting items in the most frequent categories.

Recommendation System Focus: A recommendation system might prioritize building recommendations for these popular categories

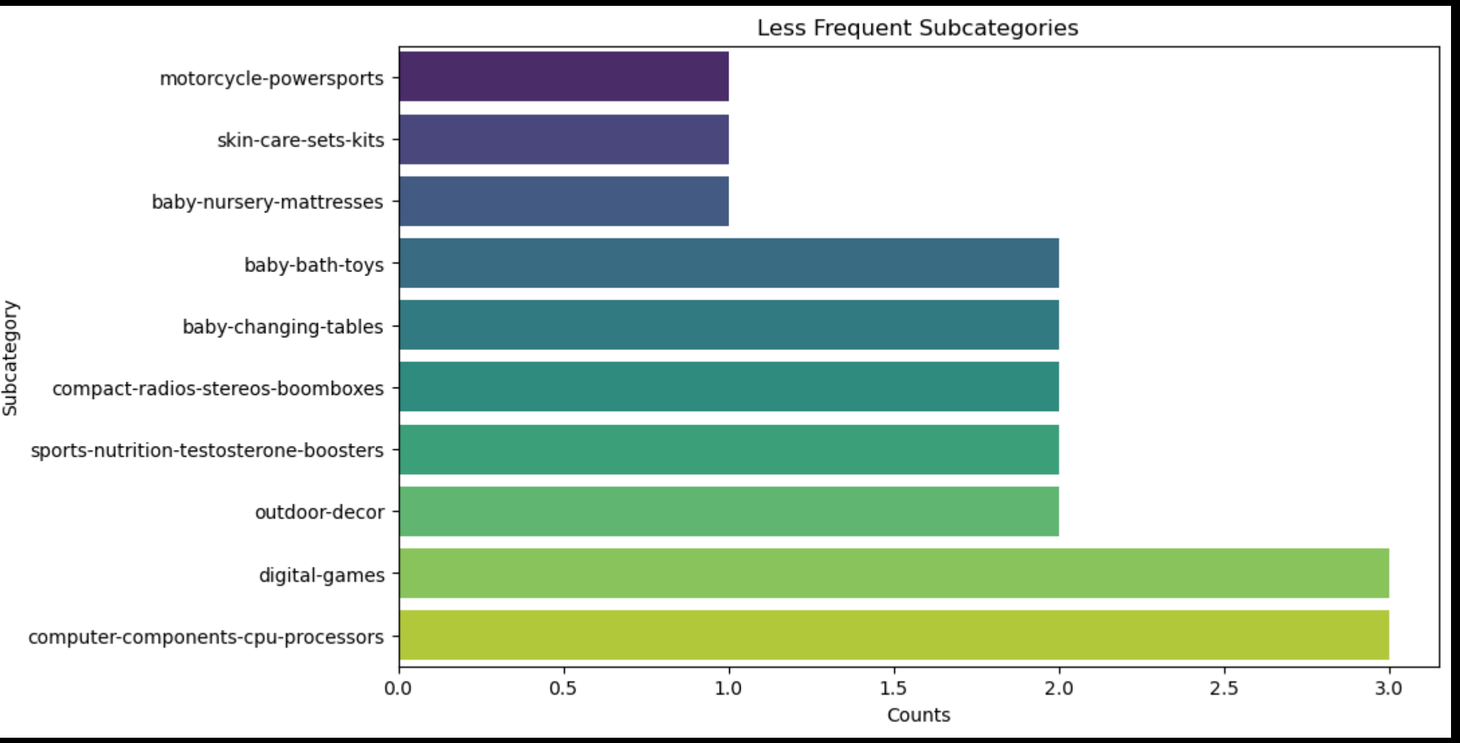
DATA ANALYSIS



SUBCATEGORIES

The top subcategories represent popular product types, high customer demand, or a focus on certain items.

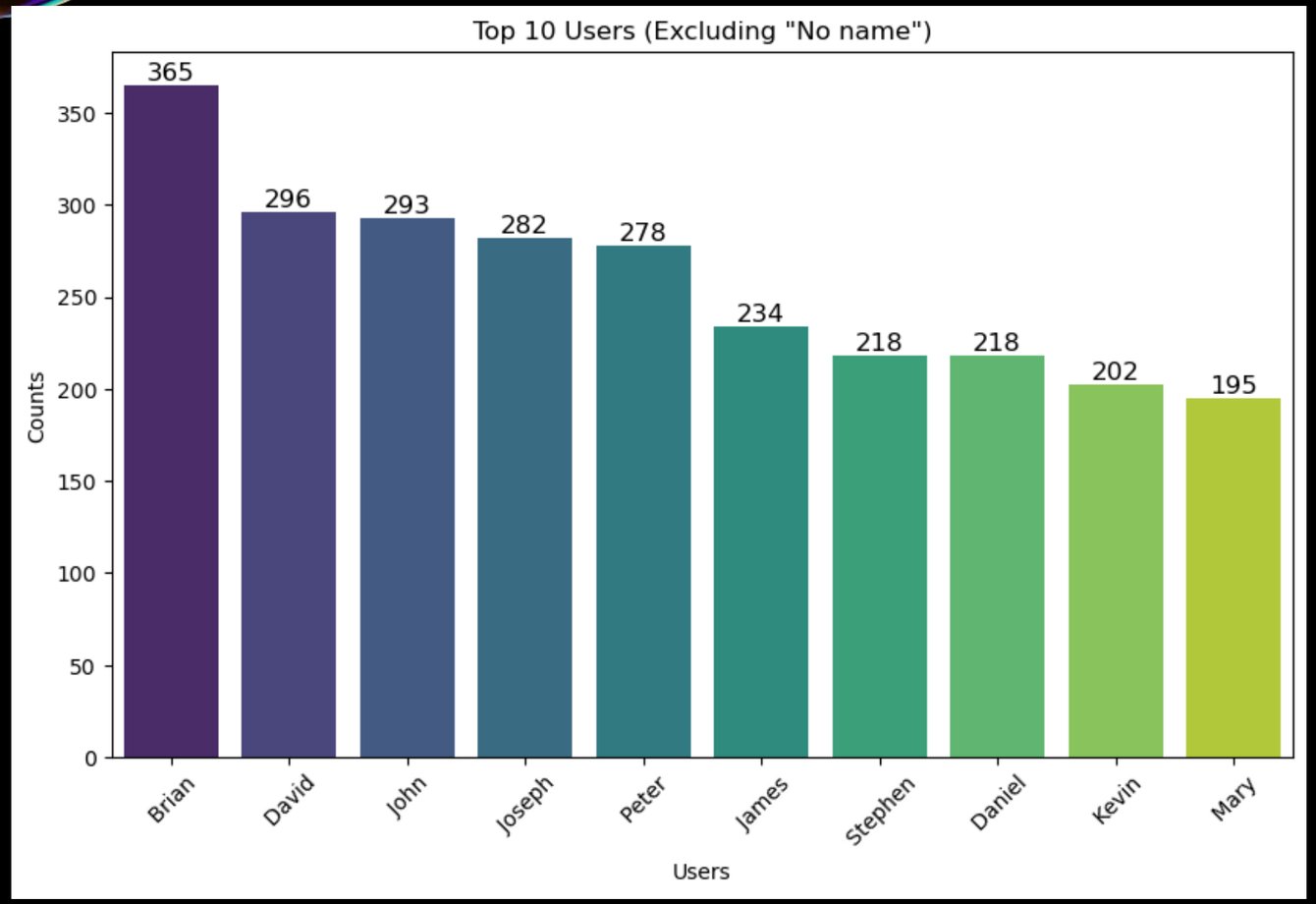
DATA ANALYSIS



SUBCATEGORIES

The plot shows less frequent sub-categories potentially indicating lower popularity or inventory for these items.

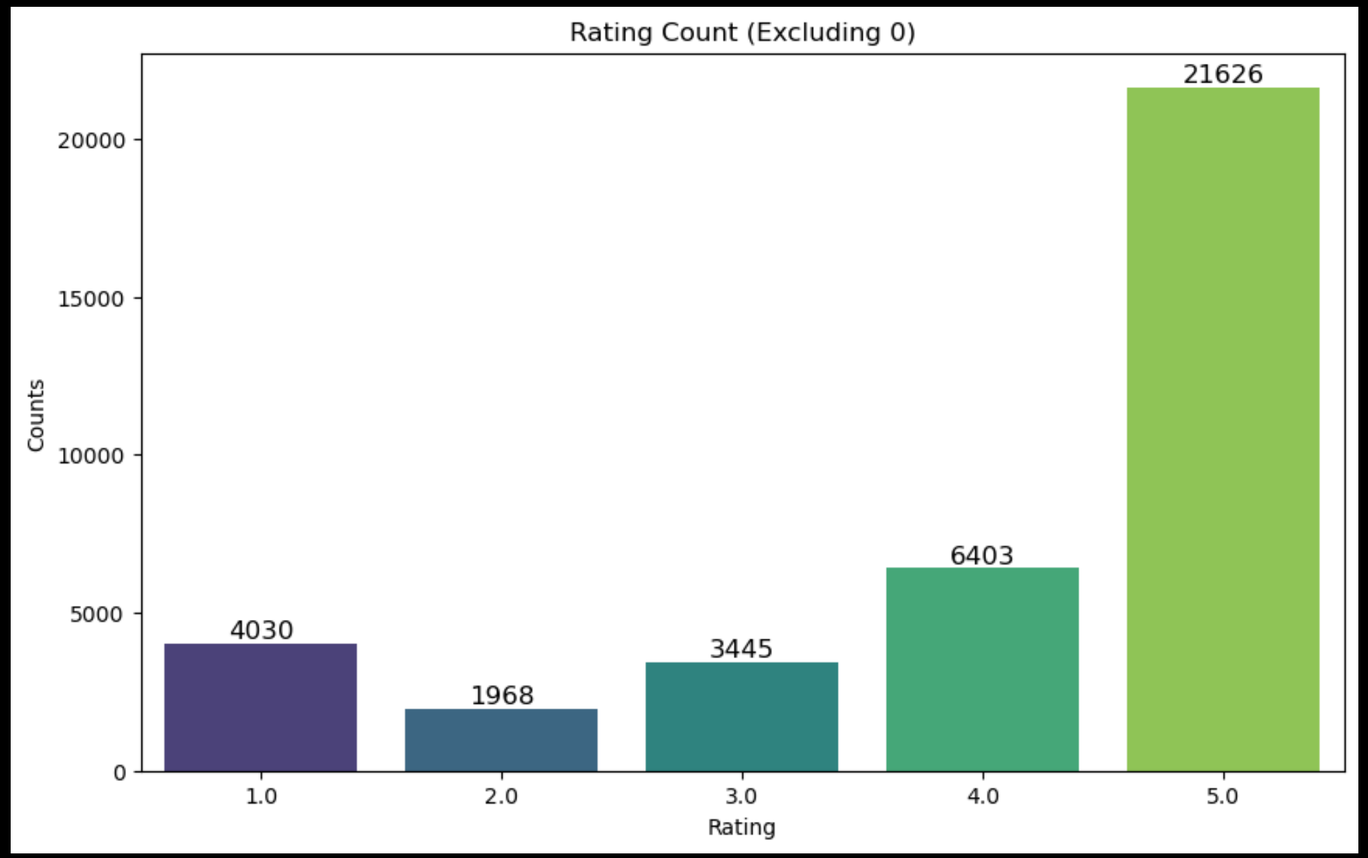
DATA ANALYSIS



USERS

This distribution suggests that a few users are significantly more engaged than others in providing feedback.
These top users are repeat customers or highly active reviewers who frequently purchase and leave feedback on products.

DATA ANALYSIS

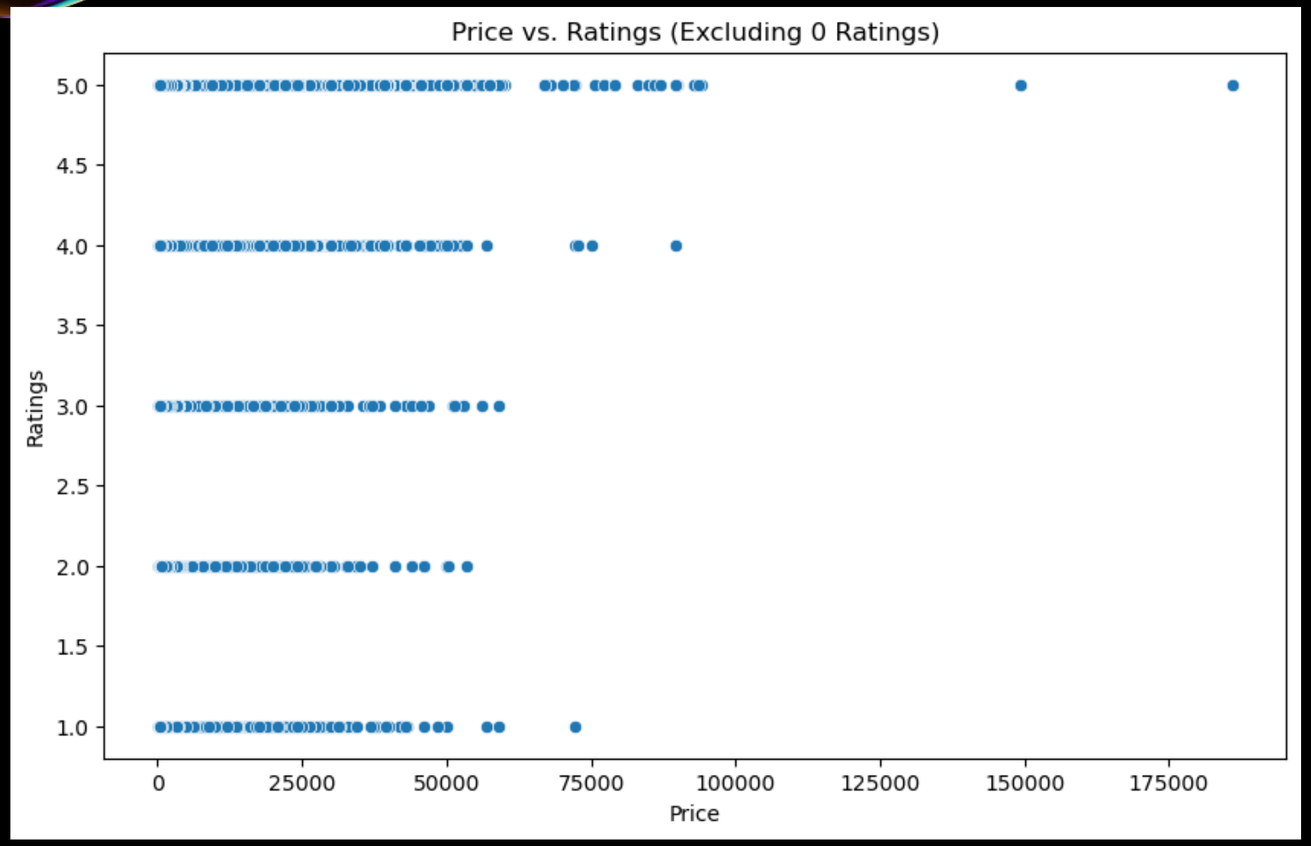


RATING COUNT

From the rating count plot, we can see that ***5*** is the most rating used by the website users followed by ***4*** which is a clear indication majority of the products are satisfying and pleasing to the buyers.

DATA ANALYSIS

PRICE VS RATINGS



The scatter plot shows that:

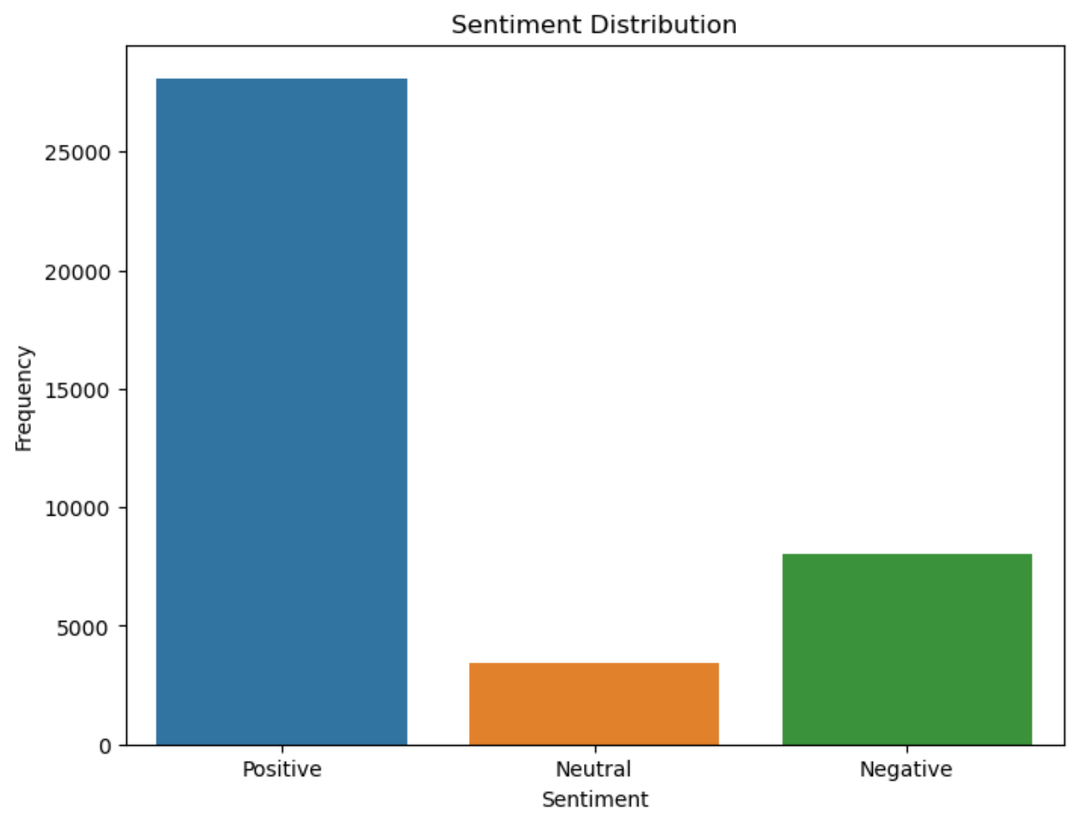
Products have a broad range of prices for each rating level.

Higher ratings (3 to 5) are more frequent than lower ratings (1 or 2).

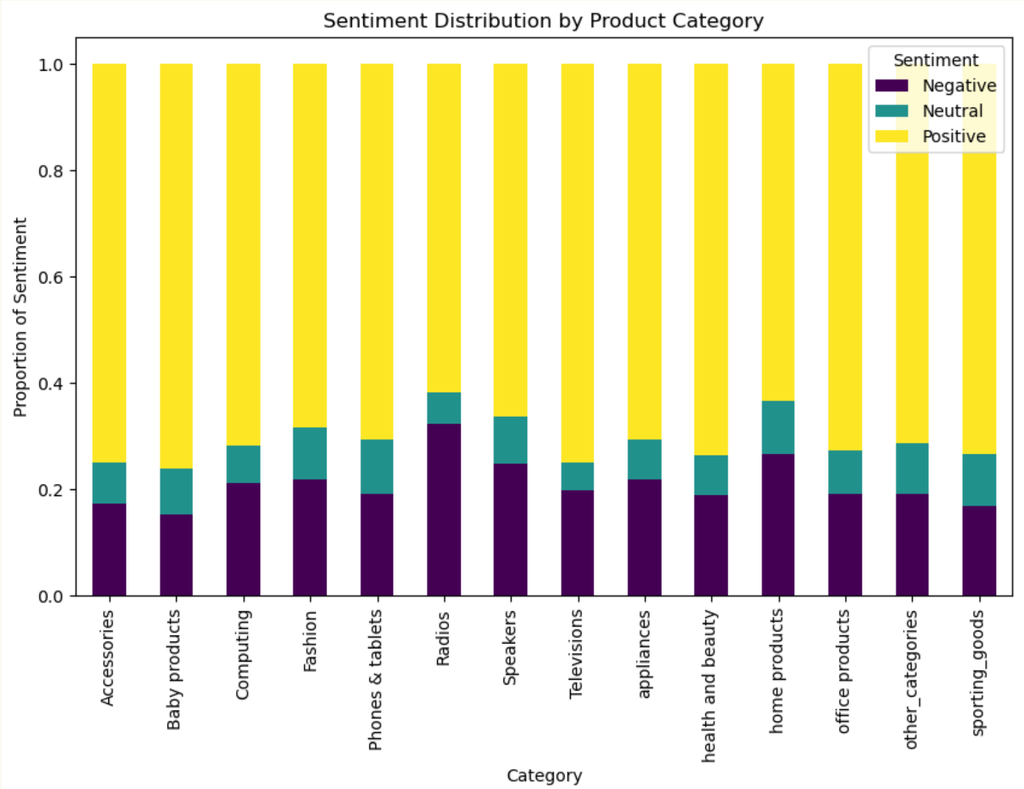
There’s no direct correlation between price and rating based on this visual.

SENTIMENT ANALYSIS

SENTIMENT DISTRIBUTION



SENTIMENT BY CATEGORY

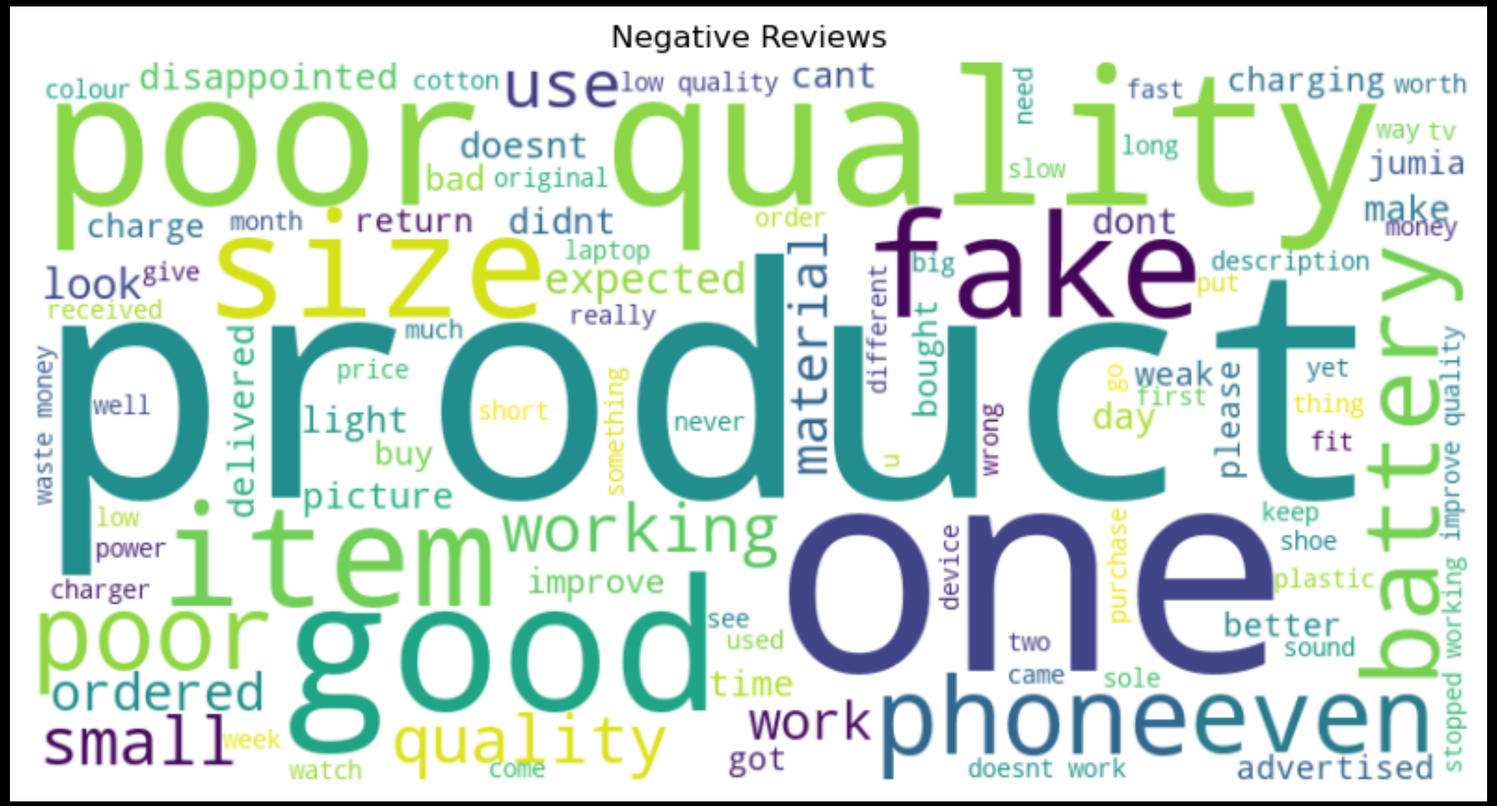


SENTIMENT ANALYSIS

POSITIVE REVIEWS

The following words represented satisfaction and high ratings from users

SENTIMENT ANALYSIS



NEGATIVE REVIEWS

The following words represented poor ratings and dissatisfaction from users.

MODELLING

BOW

	precision	recall	f1-score	support
Negative	0.70	0.57	0.63	1409
Neutral	0.48	0.30	0.37	1115
Positive	0.82	0.93	0.87	4971
accuracy			0.77	7495
macro avg	0.66	0.60	0.62	7495
weighted avg	0.74	0.77	0.75	7495

- The model performs well on Positive sentiment but struggles with Neutral sentiment, evident from the low recall (0.30) and F1-score (0.37).
- **Weighted Avg** indicates that performance metrics are influenced by the dominance of the "Positive" class, given its high support (4971 out of 7495 samples).

MODELLING

	precision	recall	f1-score	support
Negative	0.72	0.55	0.62	1504
Neutral	0.49	0.27	0.34	1282
Positive	0.78	0.94	0.85	4709
accuracy			0.75	7495
macro avg	0.66	0.58	0.61	7495
weighted avg	0.72	0.75	0.72	7495

TF-IDF

For this model,

- **Both ‘Positive’ and ‘Negative’ sentiment scores remain strong, with relatively high precision and recall.**
- **Positive sentiment still dominates, with a recall of 0.94, showing an imbalance where the model is better at identifying the majority class.**

MODELLING

NAIVE BAYES

	precision	recall	f1-score	support
Negative	0.69	0.52	0.59	1150
Neutral	0.47	0.11	0.18	690
Positive	0.85	0.97	0.90	5655
accuracy			0.82	7495
macro avg	0.67	0.53	0.56	7495
weighted avg	0.79	0.82	0.79	7495

- Naive Bayes achieves a higher accuracy (0.82) compared to earlier models (0.75–0.77).
- This model assumes feature independence, which works well for distinct patterns (e.g., clear Positive and Negative keywords).

MODELLING

	precision	recall	f1-score	support
Negative	0.75	0.62	0.68	1150
Neutral	0.78	0.28	0.41	690
Positive	0.87	0.97	0.92	5655
accuracy			0.85	7495
macro avg	0.80	0.62	0.67	7495
weighted avg	0.85	0.85	0.84	7495

RANDOM FOREST

The random forest classifier performs well for "Positive" and "Negative" sentiments i.e.,

- For ‘Positive’ sentiments, both precision (0.87) and recall (0.97) are high, resulting in an excellent F1-score (0.92).
- For ‘Negative’ sentiments, precision and recall are fairly balanced, resulting in a decent F1-score of 0.68.

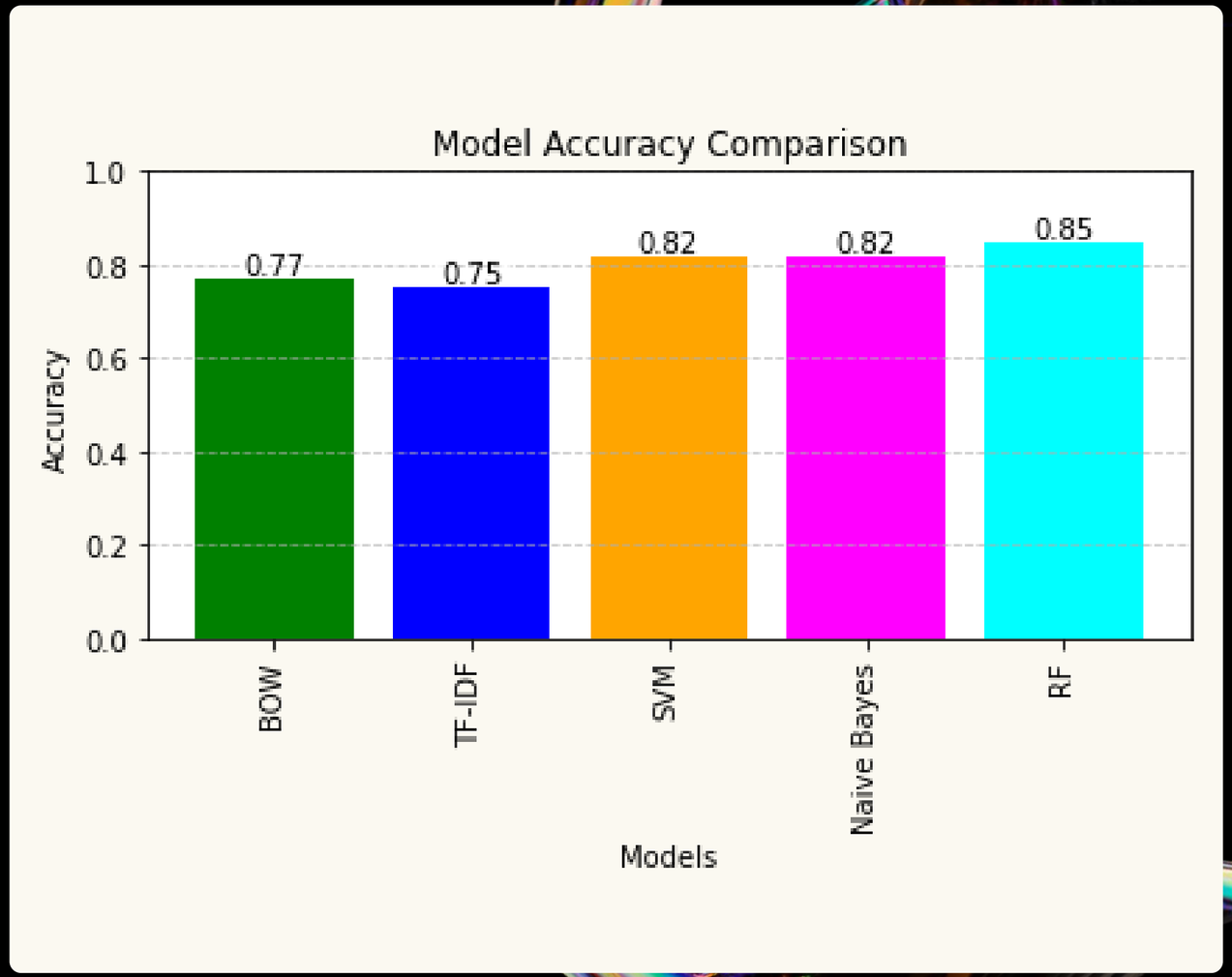
MODEL ACCURACY COMPARISON

* More sophisticated models (RF, SVM, Naive Bayes) outperform simpler approaches (BOW, TF-IDF)

* Random Forest leads with 85% accuracy, suggesting it's best at capturing complex sentiment patterns

* The performance difference between SVM and Naive Bayes is negligible (both 82%)

* All models achieve above 75% accuracy, indicating good baseline performance for sentiment analysis



RECOMMENDATION SYSTEM

TYPES OF SYSTEMS

For our recommendation system, we came up with three types of recommendation:

1.Unpersonalized Recommendation

For this, we recommended items and categories with the highest weighted ratings.

2. Item-Based Recommendation

Recommends items similar to those the user has liked in the past.

3. Content-Based Filtering

Uses item metadata (e.g., product descriptions, tags) to recommend similar items to those the user has previously liked.

RESULTS

1. **Sentiment Analysis**
   - *Best Model*: Random Forest achieved the highest performance across evaluation metrics, showcasing its effectiveness in classifying customer sentiments.
2. **Recommendation Systems**
   - *Unpersonalized System*: Delivered general recommendations based on popular and highly rated items, effective for broad user bases.
   - *Item-Based Collaborative Filtering*: Successfully suggested products by analyzing similarities between items that users interacted with.
   - *Content-Based Filtering*: Accurately provided personalized recommendations by analyzing product attributes and user preferences.

CHALLENGES

1. **Dataset Limitations**: Scraping the dataset revealed a restriction of 10 reviews per item, limiting the depth of data available for analysis.
2. **User Identification Issues:** The *user_name* field primarily contained first names, making it challenging to build a robust user-based recommendation system due to the lack of unique identifiers.
   - ***Impact:*** These challenges required adapting our approach, such as focusing on item-based and content-based recommendation systems and leveraging available data effectively.

CONCLUSION

This project successfully developed and evaluated a dual-purpose system integrating product recommendation and sentiment analysis to address key challenges in Kenya's e-commerce landscape. By focusing on user interaction data and review analysis, the system provides actionable insights that enhance user satisfaction and retailer efficiency.