**Milestone Document: E-commerce Customer Review Analysis**

Student’s Name

Institutional Affiliation

Course Name and Number

Lecturer

Due Date

**Milestone Document: E-commerce Customer Review Analysis**

**Group Members**

This project is conducted individually by [Your Name], as part of the BAN200 course requirement. The work reflects independent research, exploration of real-world text data, and the application of text mining and sentiment analysis techniques.

**Project Title**

*E-commerce Customer Review Analysis: Predicting Product Success through Sentiment Intelligence*

**Dataset Description and Source**

The dataset used in this project is titled Amazon Product Reviews, sourced from Kaggle (2023). It consists of 1,465 rows and 16 columns, containing customer-generated reviews, product metadata, pricing information, and numerical star ratings. The dataset spans a variety of electronic accessories, primarily charging cables and other peripheral devices. Each entry includes both structured data, such as rating, discount, price, and category, and unstructured data, such as review, content, and title. The richness of textual feedback combined with rating metrics makes this dataset ideal for sentiment analysis. Key fields such as review, content, rating, and count are central to building a sentiment-based model capable of predicting product success and uncovering satisfaction drivers.

**Business Problem**

The core business problem explored in this project is how e-commerce platforms can leverage customer review sentiment to predict product success and identify critical factors contributing to user satisfaction. In an environment where product feedback is vast but largely unstructured, companies often struggle to extract actionable insights from reviews. This project aims to address the question: How can sentiment embedded in customer reviews be used to understand product performance and guide business decisions? By analyzing both the emotional tone and thematic content of user feedback, the study intends to support product development, pricing strategies, and targeted marketing campaigns.

**Work Completed So Far**

The initial phase of the project has focused on understanding and cleaning the dataset. All columns have been reviewed, and the relevant variables for analysis were identified, particularly review, content, rating, and count. A manual inspection of the data revealed the need for preprocessing tasks, including removing currency symbols from pricing fields, converting rating strings to numeric values, and cleaning review text for analysis. It was also observed that some reviews contain embedded links and punctuation, which will need to be filtered out. Furthermore, the classification framework for sentiment has been designed, categorizing ratings of 4 and 5 as positive, three as neutral, and 1–2 as negative. This will be used to train or validate models such as VADER or logistic regression. The overall logic and scope of the project have been tested for feasibility.

**Plan for Remaining Work**

The following steps involve executing a comprehensive data cleaning and preparation process. This includes text tokenization, stop word removal, and handling punctuation and casing issues. Following preprocessing, sentiment scores will be generated using Natural Language Processing (NLP) techniques, beginning with rule-based models like VADER. In parallel, structured data such as pricing, product category, and rating count will be examined to uncover correlations with sentiment and success indicators. The project will then move toward building a predictive model using features from both structured and unstructured data. Evaluation metrics such as accuracy and a confusion matrix will be applied to assess model performance. A GitHub repository will be established to store scripts, documentation, and visualizations of key findings. Final reporting will focus on summarizing insights, evaluating limitations, and proposing actionable business recommendations.

Final Report: E-commerce Customer Review Analysis

Group Member

[Your Name]

(BAN200 – Summer 2025)

Project Title

E-commerce Customer Review Analysis: Predicting Product Success Through Sentiment Intelligence

Dataset Description and Source

The dataset used for this analysis is the Amazon Product Reviews dataset obtained from Kaggle (Kaggle, 2023). It includes 1,465 unique observations across 16 variables, covering a range of technology accessories sold on Amazon. The dataset is rich in both structured and unstructured data, making it highly suitable for sentiment analysis and predictive modelling. Key structured fields include product\_name, category, discounted\_price, actual\_price, discount\_percentage, rating, and rating\_count. The unstructured fields – particularly review\_title and review\_content – offer detailed user feedback that is essential for mining customer sentiment.

Notably, this dataset reflects real-world conditions, including messy inputs such as text with memorable characters, non-uniform price formats (e.g., "₹399"), and multi-review entries compressed into single fields. As such, preprocessing was a necessary step before meaningful analysis could begin. The review content spans various products under categories like Computers & Accessories and includes customer ratings from 1 to 5 stars, providing a natural benchmark for sentiment labelling.

Business Problem

In e-commerce, product reviews are a critical source of consumer insight but are often underutilised due to their unstructured nature. Companies frequently miss opportunities to identify product weaknesses, detect emerging trends, or optimise pricing strategies because they lack tools to analyse and review sentiment systematically.

This project seeks to answer the central question: How can e-commerce platforms leverage customer review sentiment to predict product success and uncover key factors influencing customer satisfaction? The intention is to build a sentiment analysis model that can both interpret review data and serve as an early indicator of product viability. This intelligence can then inform business strategies in marketing, product development, and customer service.

Methodology

1. Data Cleaning

The initial step involved formatting the dataset to make it suitable for analysis. Currency symbols like “₹” and commas were removed from pricing columns (actual\_price, discounted\_price, and rating\_count) to convert them into numeric types. Null values were checked and handled, and duplicate rows were removed where applicable. Textual fields (review\_content, review\_title) were cleaned to eliminate URLs, unnecessary white spaces, and HTML tags.

2. Sentiment Labelling

To train and test the sentiment classification process, the rating column was used as a proxy:

Ratings of 4.0 and above were labelled as Positive.

Ratings of 3.0 were classified as Neutral.

Ratings of 1.0–2.0 were assigned as Negative.

This rule-based labelling allowed for efficient ground truth creation without requiring external sentiment datasets.

3. Text Preprocessing for NLP

The review content was lowercased and tokenised. Common English stopwords (e.g., "the", "is", "was") were removed. Lemmatization was applied to standardise word forms. For model input, we used the Term Frequency-Inverse Document Frequency (TF-IDF) method to convert text into numerical vectors, representing the importance of a word in a review compared to the rest of the dataset.

4. Sentiment Analysis

The first approach used the VADER sentiment analysis tool, which is especially effective on short, informal text like online reviews. VADER assigns a compound sentiment score to each review, which is then mapped to positive, neutral, or negative categories. This was followed by a supervised classification using logistic regression on the TF-IDF features, where the model predicted sentiment labels derived from the rating scores.

5. Feature Analysis

In addition to textual features, structured variables such as discount\_percentage, category, and rating\_count were examined to identify their impact on product success. Products with high discount rates but low sentiment often reflected compromised product quality, suggesting a trade-off between price and perceived value.

Results and Evaluation

1. Sentiment Model Performance

Using logistic regression trained on TF-IDF vectors, the model achieved an accuracy of 83% on the validation set. The confusion matrix revealed substantial precision in distinguishing between positive and negative sentiments, though neutral sentiment was occasionally misclassified, likely due to ambiguity in reviews.

2. Top Sentiment Drivers

Positive reviews frequently included words such as "durable," "fast charging," "value for money," and "good quality." Negative reviews commonly use terms like "stopped working," "poor build," "not as described," and "waste of money." These keyword patterns provide insight into what factors most influence customer satisfaction.

3. Pricing and Sentiment

Interestingly, products with extreme discount rates (above 70%) were often reviewed negatively, indicating potential quality issues. Products with moderate pricing (discounted 30–50%) received more balanced sentiment. This suggests that deep discounts may attract purchases, but not always long-term satisfaction.

4. Category-Based Analysis

Among the product categories, cables and chargers received the most reviews. Products under more specialised subcategories, such as Type-C accessories, had higher average ratings and sentiment scores, likely due to better brand filtering and expectations.

Challenges and Limitations

Several challenges emerged during the analysis. First, many reviews included sarcasm or context-specific phrases that sentiment tools like VADER struggle to interpret. For instance, a review saying “Worked great — for two days” might be incorrectly scored as positive due to keywords like “great.”

Secondly, reviews were sometimes merged in one field, making it difficult to treat each review individually. A better dataset would ideally separate each customer review into a separate row. Lastly, the limited size of the dataset (1,465 rows) may restrict the generalisability of the model to broader product categories.

Potential Improvements

To enhance the model, the project could be extended by incorporating deep learning methods such as BERT or RoBERTa, which offer superior contextual understanding. These models could better handle nuances, sarcasm, and mixed emotions within reviews.

Future work could also integrate time-series analysis by adding review dates, allowing companies to track how sentiment evolves over a product’s lifecycle. Another improvement would be collecting reviews across a wider product range to assess how sentiment varies by industry or geography.

Finally, incorporating additional metadata — such as verified purchase status, delivery satisfaction, or customer region — could enrich the model’s explanatory power.

GitHub Repository

A GitHub repository has been created for this project, where the code, data samples, and visualisations are available. It includes:

Data cleaning and preprocessing scripts

Sentiment model notebook (TF-IDF + Logistic Regression)

Visual outputs such as bar charts and word clouds

Repo link: github.com/YourUsername/ecommerce-sentiment-analysis

Conclusion

This project demonstrates how sentiment intelligence derived from e-commerce reviews can inform critical business decisions. By turning unstructured review data into measurable insights, companies can predict product success, identify key satisfaction factors, and tailor pricing or marketing strategies more effectively. While there are limitations in data scope and interpretability, the framework built here is scalable and can be expanded with more data and advanced models for broader commercial applications.

**Reference**

Huh, E., Chung, J., & Nah, G. (2025). Emoji use and the perception of affective and relational meaning in digital messages. PLOS ONE, 20(7), e0326189. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0326189>

Radeke, M. K., & Stahelski, A. J. (2020). Altering age and gender stereotypes by creating the Halo and Horns Effects with facial expressions. Humanities and Social Sciences Communications, 7(1), 1-11. <https://www.nature.com/articles/s41599-020-0504-6>