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

# Milestone Report

**Dataset Description and Source:**

This project employs the Amazon Sales Dataset (Karkavelraja, 2023), a comprehensive collection of over 1,000 product reviews from Amazon's e-commerce platform. The data is freely accessible on Kaggle and contains 1,465 records with 16 variables each. It indulges in various product segments, such as electronics, personal accessories, and home goods. The most important variables are product\_id, product\_name, category, discounted\_price, actual\_price, rating, rating\_count, review\_title, and review\_content.

The availability of structured numeric values (such as price, rating, and discounts) and unstructured text values (such as user reviews) makes the data set especially good for applying text mining, sentiment analysis, and predictive modeling. It helps us draw information regarding customer perception and satisfaction with products, which can help make strategic decisions related to pricing, marketing, and product development.

**Business Problem or Question**

In online retail, customer feedback is an untapped reservoir of insight. However, analyzing thousands of reviews manually is impractical. The key question driving this project is:

How can companies use sentiment in customer reviews to predict product success and uncover actionable insights for improvement?

This analysis aims to help businesses better understand how products are perceived, whether pricing (especially discounting) affects customer sentiment, and what textual indicators in reviews correlate with high or low ratings. Ultimately, the goal is to demonstrate how natural language processing (NLP) can be integrated into e-commerce analytics for more intelligent decision-making.

**Overview of Work Done So Far**

The project has progressed through its exploratory and preprocessing phases. The dataset was loaded into a Google Colab notebook, and key variables were cleaned and formatted using Python and pandas. Prices were stripped of currency symbols and commas for numerical comparison. Null values in the rating\_count column were imputed with median values.

Unstructured review data from the review\_title and review\_content columns were preprocessed using NLTK, including:

- Lowercasing

- Removal of punctuation and stopwords

- Tokenization

- Lemmatization

For sentiment analysis, we used two tools:

1. TextBlob: Provided polarity and subjectivity scores for each review.

2. VADER: Delivered compound sentiment scores better suited for short, informal text like product reviews.

Initial analysis shows a moderate correlation between star ratings and sentiment scores. Interestingly, some highly discounted products still received negative reviews, suggesting that discounting alone does not improve customer satisfaction. Review-text/rating mismatches (e.g., positive language with low ratings) signals possible areas for deeper analysis—such as sarcasm or expectations mismatch.

Seaborn and Matplotlib visualizations helped uncover patterns, such as the consistently polarized sentiment in certain product categories (e.g., charging cables).

**Plan for Remaining Work:**

The next stages of this project include:

- Transformer-based Sentiment Models: Implement DistilBERT from Hugging Face for more accurate sentiment classification.

- Topic Modeling: Use Latent Dirichlet Allocation (LDA) to discover dominant themes in positive and negative reviews.

- TF-IDF Feature Engineering: Extract the most influential keywords across sentiment groups to identify what features matter most to customers.

- Predictive Modeling: Build a linear regression model to predict star ratings from sentiment scores, discount rates, and product categories.

- Insight Generation: Provide business recommendations on pricing strategies, product messaging, and design improvements.

All progress will be documented in Jupyter notebooks, and visual outputs and intermediate findings will be posted to a shared repository.

Link to GitHub Repository (in progress):

<https://github.com/Hawkins129/New-Ecommerce-sentiment-analysis/tree/main>

Reference:

Karkavelraja, J. (2023). Amazon Sales Dataset. Kaggle. https://www.kaggle.com/datasets/karkavelrajaa/amazon-sales-dataset