Forecasting and Sentiment Analysis for HUL

A Mid-Term report for the BDM capstone Project

Submitted by

Name: Nitish Rishi
Roll number: 22f3000645

IITM Online BS Degree Program,
Indian Institute of Technology, Madras, Chennai
Tamil Nadu, India, 600036

Contents

1. Executive Summary

Hindustan Unilever Limited (HUL) is India's leading FMCG company, operating across personal care, home care, and foods segments. In a complex and competitive marketplace, HUL faces interrelated business challenges: optimizing inventory through precise demand forecasting, understanding evolving consumer sentiment on its products, and ensuring production is aligned with anticipated demand. These challenges are crucial for sustaining profitability, reducing operational inefficiencies, and maintaining customer loyalty.

This mid-term report presents an integrated analysis focused on the first two pillars of the overall project: (1) segment-level demand forecasting and (2) consumer sentiment analysis. For demand forecasting, we compiled a segment-wise dataset by extracting revenue information from HUL's quarterly reports and linking it with relevant macroeconomic indicators, including Consumer Price Index (CPI), Consumer Sentiment Index (CSI), and Future Expectations Index (FEI). The resulting time series spans from 2018 onward, incorporating both internal financial trends and external economic drivers. Several modeling approaches—SARIMA, Prophet, Ridge Regression, and XGBoost—were evaluated. Ridge Regression proved most effective.

For sentiment analysis, over 2,000 customer reviews were scraped from major digital platforms, including Walmart.com, covering a variety of HUL's brand offerings. Using natural language processing (NLP) methods, notably the VADER classifier, customer feedback was systematically labeled for polarity and aggregated at the product level. This revealed a broadly positive perception of HUL products, though with variability in brand reception and discrepancies between ratings and review sentiment for certain SKUs.

2. Proof of Originality of Data

This project is built entirely on publicly available and self-extracted data sources, ensuring originality and transparency in methodology. The data collection process was done manually and programmatically through web scraping or direct downloads from government and company websites. Below is a detailed account of sources used for each problem statement.[Link to data]

2.1 Problem Statement 1: Demand Forecasting

To model demand forecasting and align production schedules, the following data sources were utilized:

- **HUL Financial Reports**: A total of 29 quarterly reports were downloaded from the official Hindustan Unilever Limited investor relations page. From these, consolidated segment-wise revenue data was extracted manually and programmatically, converted into time series format for analysis.
- **Consumer Sentiment Index (CSI)**: Data was sourced from the **Reserve Bank of India (RBI)** through their Urban Consumer Confidence Survey, available at: Link
- **Consumer Price Index (CPI)**: Inflation data and category-specific indices were downloaded from the **Ministry of Statistics and Programme Implementation (MoSPI)** at: Link

These datasets were cleaned and integrated for model building, with full citations and source links documented.

2.2 Problem Statement 2: Sentiment Analysis on HUL Products

To perform sentiment analysis, customer review data was gathered:

- **Walmart.com Product Reviews**: Over 2,000 customer reviews were scraped across 13 different HUL product listings on Walmart's website. These include key brands such as Dove, TRESemmé, Hellmann's, and Lipton . Walmart Majorly had products from the Food and Drinks, Personal Care segment.

3. Metadata & Descriptive Statistics

There are Mainly 2 datasets catering to each problem statement:

3.1 Metadata for Demand Forecasting (Problem Statement 1)

The dataset for this task was created by extracting segment-wise financial information from HUL's quarterly reports, combined with macroeconomic indicators to enhance forecasting accuracy. Each row in the dataset corresponds to a fiscal quarter (e.g., 2019Q1, 2019Q2), captured in the Quarter column — this forms the time index for modeling.

The Personal Care, Foods, Home Care, and Others columns represent revenue (in crores) generated by each product segment. These are the central variables used to forecast future demand

at the category level. Forecasting models like SARIMA or Prophet rely heavily on this historical data to capture patterns, seasonality, and trend components.

To measure performance dynamics over time, year-over-year percentage growth for each segment is captured under columns like Personal Care YoY %, Foods YoY %, etc. These help in identifying fast-growing or declining segments and inform inventory planning decisions.

The Volume Index variables provide a normalized view of sales volume compared to a base quarter, enabling cross-category comparison of demand shifts. Similarly, Chain Volume Index captures the cumulative or relative growth from quarter to quarter. These indicators are especially useful for identifying long-term structural changes in product demand .

To account for macroeconomic drivers of consumer demand, the dataset includes three external indicators:

- CPI (Consumer Price Index): Measures inflationary pressure and is crucial for understanding how price levels may suppress or encourage consumption.
- CSI (Consumer Sentiment Index): Reflects current consumer confidence and can signal shifts in buying behavior.
- FEI (Future Expectations Index): Captures consumer optimism or pessimism about the near-term economy, useful for predicting demand fluctuations.

Together, these variables form a comprehensive dataset for modeling category-level demand.

Descriptive Statistics

Based on the data from Q1 2019 to Q4 2025, the **Personal Care** segment consistently emerges as the primary revenue driver for HUL, showing the highest average quarterly revenue. While the **Home Care** segment is the second-largest contributor, it also exhibits the greatest volatility, as indicated by its high standard deviation. The macroeconomic indicators, particularly the Consumer Price Index (CPI), show significant fluctuation, suggesting a dynamic economic environment that likely influenced consumer spending patterns across all categories during this period.

	count	mean	std	min	25%	50%	75%	max
Personal Care	28.0	5001.86	600.90	3834.0	4570.25	4947.00	5613.00	5873.0
Foods	28.0	3119.64	860.76	1703.0	1941.50	3566.50	3764.75	3910.0
Home Care	28.0	4396.57	1043.67	3079.0	3404.75	4015.00	5461.50	5815.0
Others	28.0	487.79	118.19	239.0	380.00	522.00	571.75	661.0
CPI	26.0	164.28	26.54	64.2	152.30	166.30	183.98	196.2
CSI	26.0	80.41	17.17	49.8	64.79	85.80	94.20	104.6
FEI	26.0	116.33	8.24	97.6	113.04	115.88	121.51	133.4

Figure 1: Descriptive Statistic for Demand Forecast

The chart (Fig 2) illustrates the segment wise company's growth dynamics,the most significant trend is the remarkable ascent of the Home Care segment. It began as the second-largest but demonstrated aggressive and consistent growth, ultimately overtaking Personal Care around late 2022 to become the primary revenue driver.The Foods segment displays steady growth, highlighted by a distinct jump in performance around late 2020. In stark contrast, the Others segment has remained relatively flat.

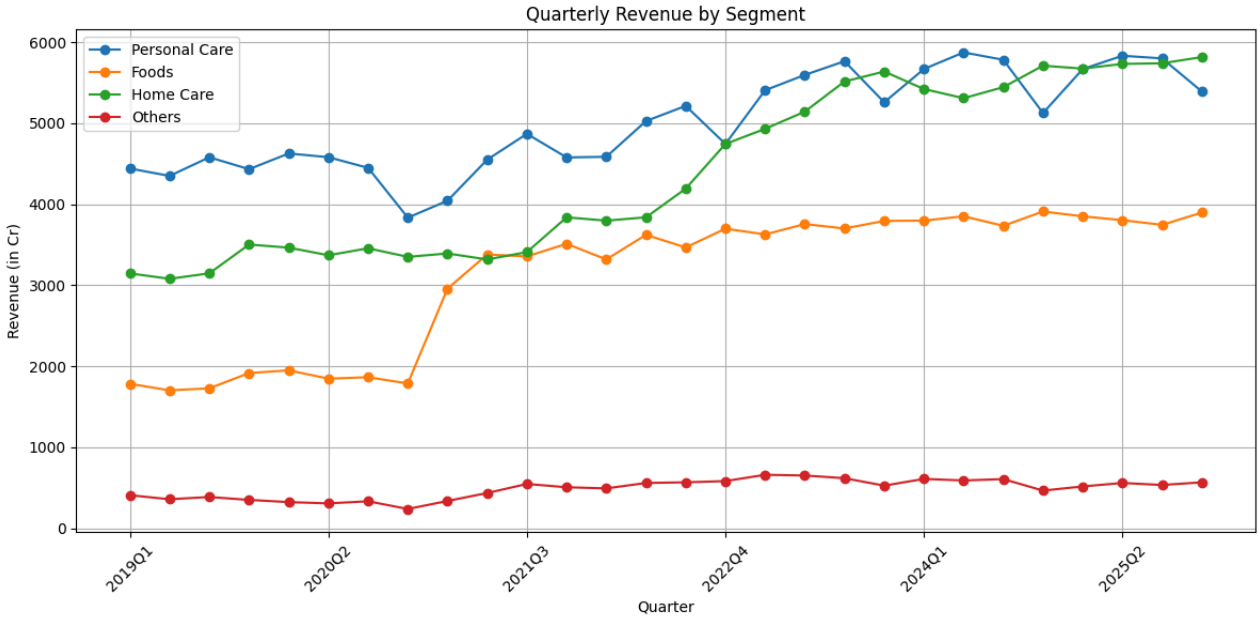

Figure 2: Revenue (in Cr) vs Quarter Segment wise

3.2 Metadata for Semantic Analysis (Problem Statement 2)

This dataset is derived from scraping over 2,000 customer reviews across 13 HUL products listed on Walmart.com. Each review is accompanied by rich metadata that enables sentiment analysis.

The product_id and product_name help associate each review with a specific HUL product, allowing brand- and SKU-level analysis of consumer perception. The review_position column tracks the order in which reviews were listed.

Text-based fields like review_title and review_text form the core input for natural language processing (NLP). These are analyzed using techniques like sentiment classification and topic modeling to extract insights about customer preferences, pain points, and recurring themes.The rating column captures the star rating (from 1 to 5) given by each reviewer. It acts as a numerical signal that reinforces or contrasts with sentiment derived from the review text.

Columns such as positive_feedback and negative_feedback indicate how helpful other users found the review. review_date allows for temporal trend analysis, showing how sentiment evolves over time. The user_nickname is an anonymized identifier for the reviewer, while the verified_purchaser column helps filter genuine product feedback from noise.

An aggregated dataset summarizes product ratings, providing the average overall_rating, the total_review_count, and a breakdown of counts for each star level from 1 to 5.

Descriptive Statistics

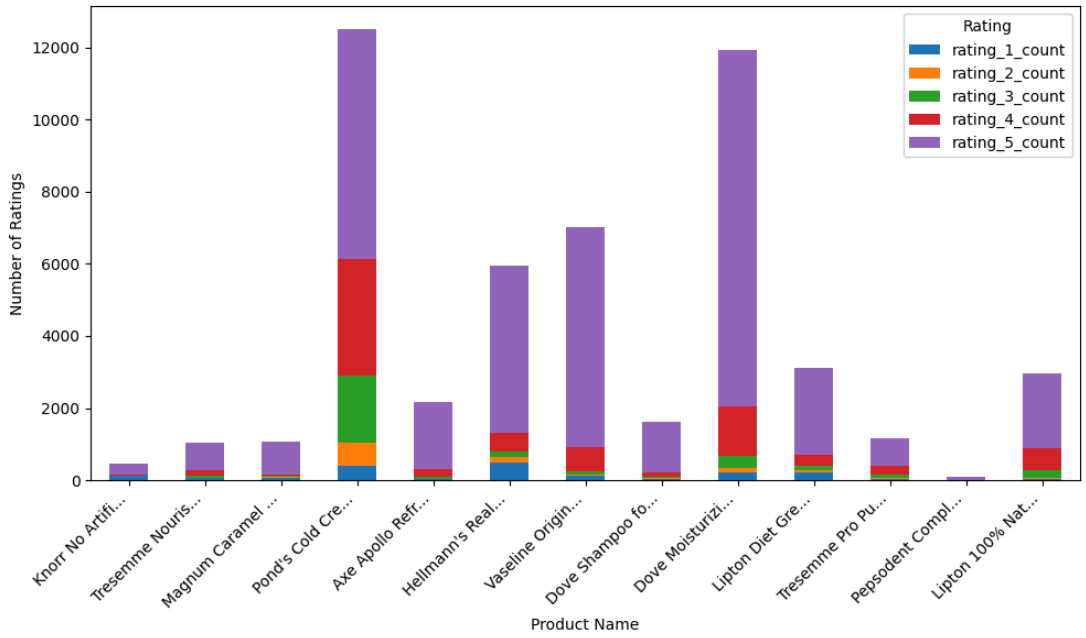

Figure 3: Rating Distribution per Product

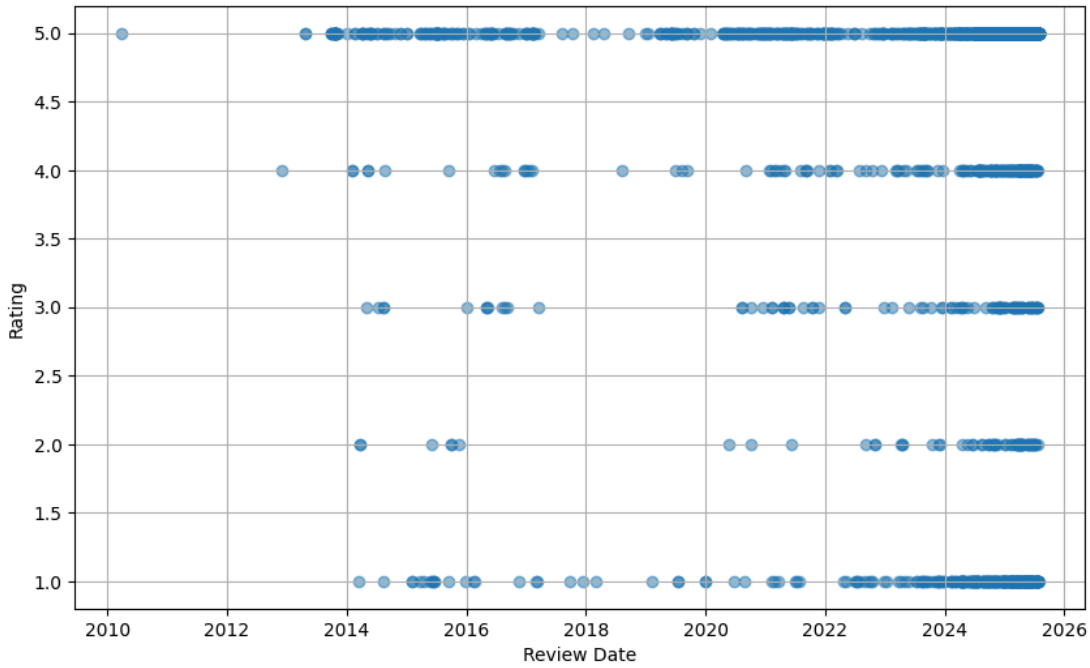

Figure 4: Ratings over Time

4. Detailed Explanation of Analysis Process & Methods

4.1 Data Cleaning & Preprocessing

For the Demand Forecasting dataset, the first step was to handle missing values in the year-over-year (YoY) change columns. The YoY values for 2018 were not calculated because data for the corresponding quarters in the previous year had not been scrapped. These rows were removed to ensure that all YoY calculations were based on valid historical data.

Segment-wise quarterly revenue data for HUL (Personal Care, Foods, Home Care, Others) was combined with macroeconomic indicators — Consumer Price Index (CPI), Consumer Sentiment Index (CSI), and Future Expectations Index (FEI) — using the quarter as the join key. Since the macroeconomic data was originally available on a monthly basis, a quarterly moving average was applied to align it with the financial reporting period.

Finally, the cleaned segment-level dataset was merged with the processed macroeconomic indicators, ensuring a consistent quarterly structure with no missing values. This prepared dataset served as the foundation for subsequent time-series forecasting and analysis.

For Sentiment Analysis, the review_date field was converted into a consistent datetime format to ensure compatibility with any time-based trend analysis. This allowed for potential grouping and filtering of reviews by quarter, month, or year.Basic text preprocessing steps, such as stopword removal and lowercasing, were applied to the review text to standardize content for analysis.

Both datasets were inspected for missing or inconsistent values, but since the data was scraped in a structured format directly from the product pages, minimal cleaning was required. The two datasets were retained in their original structures, ensuring that identifiers like product_id and product_name could be used for accurate merging and mapping between review-level sentiment trends and product-level rating summaries.

4.2.1 Analysis Method - Demand Forecasting (Problem Statement 1)

Computation of Year-over-Year (YoY) Growth

For each product segment (Personal Care, Foods, Home Care, Others), the YoY % change was calculated as:

$$YoY\ \% \ = \ \frac{Revenue_t - Revenue_{t-4}}{Revenue_{t-4}} \times 100$$

where t represents the current quarter, and t - 4 represents the same quarter in the previous year. YoY growth removes seasonality effects inherent in quarterly FMCG sales, allowing for a clearer comparison of performance across years. This approach is more stable than comparing consecutive quarters, which can be distorted by seasonal demand.

Calculation of Volume Index

A normalized **Volume Index** was computed for each segment using **2018Q1 as the base period**:

$$Volume\ Index_t = \ \frac{Revenue_t}{Revenue_{base}} \times 100$$

The index standardizes revenue across segments, enabling direct comparisons even when absolute sales volumes differ significantly. Choosing a fixed base quarter provides a consistent reference point for the entire analysis.

Calculation of Chain Volume Index

The **Chain Volume Index** was calculated by compounding YoY % changes starting from the base quarter (2018Q1):

$$Chain\ Volume\ Index_t = \ Chain\ Volume\ Index_{t-1} \times \ (1 + \frac{YoY\ \%_t}{100})$$

This metric captures cumulative growth trends and reflects long-term performance momentum more effectively than standalone YoY values.

Forecasting Approach

The initial plan was to experiment with both statistical time-series models and machine learning methods. SARIMA (Seasonal AutoRegressive Integrated Moving Average) and Prophet were first considered because they are widely used for forecasting problems involving seasonality and trend detection. However, these models underperformed in this case due to the nature of the dataset:

- The data was quarterly and limited in historical length, reducing the ability of SARIMA and Prophet to detect strong seasonal patterns.
- The fluctuations in demand were more irregular and driven by macroeconomic indicators rather than recurring seasonal trends.
- Both SARIMA and Prophet in their standard form are univariate, relying only on the target's past values, whereas this problem benefited from multiple external predictors such as CPI, CSI, FEI, and derived metrics like YoY % change and Chain Volume Index.

Given the small size of the dataset and the complexity of the relationships, we explored machine learning models capable of handling multivariate time series data. Ridge Regression was used as a

baseline linear model and demonstrated strong predictive performance, especially in capturing linear dependencies and lagged effects in the data.

XGBoost Regressor was also tested to capture potential non-linear relationships. However, despite its flexibility and power, XGBoost did not outperform Ridge in this context, likely due to the limited data volume and potential overfitting risks. XGBoost showed higher forecast errors and poorer goodness-of-fit metrics across most product categories.

Therefore, Ridge Regression was selected as the preferred model for this analysis, effectively leveraging both historical lag features and macroeconomic indicators to provide reliable and interpretable forecasts across product segments.

4.2.2 Analysis Method - Semantic Analysis (Problem Statement 2)

Data Loading
Product review data and product ratings summaries were loaded as Pandas DataFrames from CSV files. The review-level dataset contains individual customer feedback, while the ratings summary holds aggregated product ratings and counts.

Text Cleaning
Two versions of each review were generated:

- For sentiment analysis, minimal preprocessing was performed—removal extra spaces—preserving original customer phrasing for more accurate polarity detection.
- For topic modeling, reviews underwent deeper cleaning: lowercasing, punctuation and digit removal, stopword removal, and lemmatization (using NLTK tools). This focused the topic model on identifying meaningful concepts rather than noise.

Sentiment Classification
VADER was applied to the cleaned review text. For each review, the compound score was calculated and mapped to a sentiment label:

- Positive if compound score ≥ 0.05
- Negative if ≤ -0.05
- Neutral otherwise

This rule-based approach is well-suited for informal user reviews and avoids the need for extensive training data.

Product-level Sentiment Aggregation and Comparison with Ratings
Post-classification, sentiment labels were aggregated by product. For each product, the percentage of positive reviews was calculated. This metric was merged with ratings summary data (overall rating and star counts), allowing direct comparison between text-based sentiment and average ratings. This step highlighted discrepancies such as products with strong ratings despite less enthusiastic written feedback.

5. Results & Findings

The analysis reveals that the **Ridge Regression model delivered substantially more accurate forecasts** compared to the XGBoost model across the tested segments. This conclusion is

supported by both quantitative error metrics (MAE, RMSE, and R^2) and the visual evidence from the forecast plots. While both models performed comparably in the "Personal Care" segment, Ridge Regression's superiority was particularly pronounced in the "Foods" and "Home Care" segments.

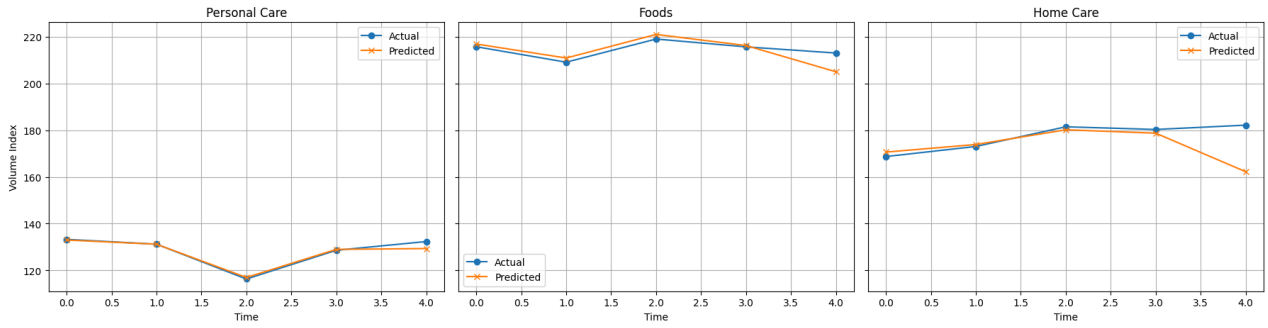

Figure 5: Actual vs Predicted - Ridge Regression

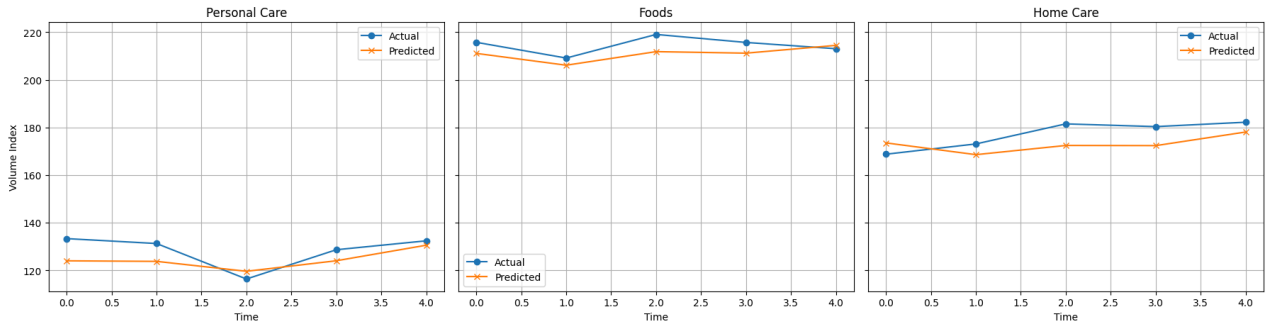

Figure 6: Actual vs Predicted - XGBoost Regression

Customer review analysis, (Fig 3) showed Pond's cream and Dove moisturizer having a very high number of reviews thus suggesting high popularity whereas Pepsodent had the lowest number of reviews thus suggesting low popularity.

The sentiment analysis, conducted using VADER, reveals an overwhelmingly positive customer reception for the products. As illustrated in the chart, the volume of **positive** reviews significantly surpasses the combined count of **neutral** and **negative** feedback. This heavily skewed distribution suggests a high degree of overall customer satisfaction. This also aligns with the overall rating Average of **4.51**. Further Analysis will be included in the Final term report which will contain topic modelling analysis.

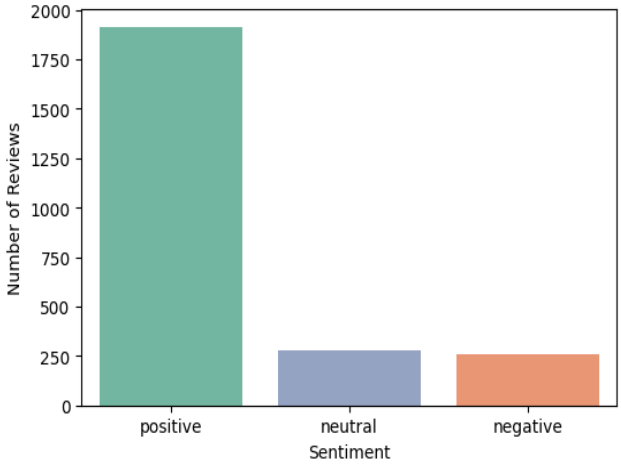

Figure 7: Overall Sentiment Distribution