ALBUKHARY INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTING AND INFORMATICS

NATURAL LANGUAGE PROCESSING

STUDENT'S NAME	STUDENT ID
Mutasim Billah	AIU21102122

Table of Content

1.0 Introduction

Opinion mining is also known as sentiment analysis – the particular branch of natural language processing that is aimed at the detection, classification, and quantification of the subjective data from the textual data. It is the most employed to analyse people's feelings, perceptions, and opinions towards various entities ranging from the market, politics, and customer service management. Thus, in the sphere of e-commerce, it can be employed to improve its tends, evaluate the level of its customers' satisfaction, as well as to make certain relevant decisions based on the results of such analysis.

In this project, critical assessment of customer reviews from Flipkart, Indian's fastest growing online shopping destination has been of interest. The collected data is the customers' feedback gathered in the last five years, which makes it a modern and varied sample. The task of this review is to divide all the reviews according to their sentiment into positive, negative, and neutral and reveal the patterns in a customer's opinion that would be beneficial for the company's strategic developments. This sentiment analysis task is highly important in e-commerce platforms such as Flipkart since customer satisfaction is a vital tenet of sustaining competitive advantage and enduring loyalty.

The proposed work only focuses on the textual data that are obtained from product review pages of Flipkart website and is mainly based on applying machine learning models and natural language processing techniques. These entails filtering this data for noise, feature extraction to make the text as amenable to machine learning as possible and the use of classification schemes to analyze the resulting sentiment labels.

By doing so we hope to educate other people about the uses of sentiment analysis in social networks to deliver important information about customer satisfaction levels and also to create effective marketing strategies. Furthermore, the conclusion drawn of this analyse will also demonstrate the applicability of tool such as NLP and Machine learning in solving real life business issue.

The key objectives of this project are as follows:

- **Sentiment Categorization:** To categorize each customer into positive, negative or neutral so that a clear picture of the trends of customers responses can be easily understood.

- **Model Comparison**: To apply more than one method of machine learning for the classification of sentiments and to compare and contrast the merits and the demerit of each model.

- **Visualization and Interpretation:** For the analysis of sentiment distribution, sentiment key terms and trends, different graphical representation like pie diagram, word cloud and scatter diagram are used.

- **Actionable Insights:** To get the list of change recommendations that might be made based on the sentiment types and patterns found out in the data.

2.0 Data Collection & Task Description

2.1 Data Collection

In this project, the datasets used are the customer reviews collected from the public platform of Flipkart e-commerce site. All these reviews gathered from the last five years to provide a broad variety of customer feedback. First, the datasets contain textual opinions/reviews as the main source of information, along with additional contextual metadata including rating, product categories and timestamps.

To ensure high-quality data, the following preprocessing steps were performed:

1. **Duplicate Removal:** Reviews were further filtered to remove any duplicates in order to avoid overcrowding with similar opinions.
2. **Spam Filtering:** To make the dataset cleaner of reviews, the ones that had contents that did not relate to the product or were explicit spam contents were removed.
3. **Language Filtering:** All non-English reviews were excluded from the study since the present work is aimed at recognition of the English language texts.
4. **Noise Reduction:** Some reviews containing special characters and some containing very little text were removed so that quality data is available for analysis.

2.2 Task Description

The primary task of this project is to perform sentiment analysis on Flipkart reviews, classifying each review into one of the following categories:

- **Positive:** Any indicators that can portray the picture of what a customer thinks about the product or even post- purchase experiences that are positive.
- **Negative:** Negative feedback in the form of complaints or negative feedback received on the site.
- **Neutral:** Not so positive, not to the extent of negative feedbacks, but instead offer actual or descriptive feedback about the product.

The project also aims to achieve the following:

1. Design a model that has incorporated at least two models for sentiment classification: **Logistic Regression** and **Random Forest**.
2. Discover and prepare data for the further feature extraction in order to train a model.
3. Illustrate how sentiment distribution can be represented, and how certain critical aspects may be inferred from the results employing suitable graphics.
4. Evaluate the models selected to create the study and then analyze the outcomes critically.

2.3 Objectives and Scope

The objectives of this analysis are closely aligned with the rubric and include:

- Classification of sentiment polarity in the context of customer's reviews.
- An attempt to construct an analytical model by employing sophisticated Natural Language Processing approaches and machine learning algorithms.
- Drawing out patterns and trends of customers' sentiment to discover key learning opportunities.
- Offering comprehensive descriptions and a comparison of models' performance and its indices are **precision, recall, accuracy, and F1-score**.

3.0 Framework Design

Regarding the framework for sentiment analysis, the design of this architecture is aimed to process and categorize vernacular textual data into sentiment categories about specific issues by employing NLP mechanisms and —ultimately— ML. The design ensures reliability and validity of classification as well as the results to correspond to the rubric criteria.

3.1 Data Preprocessing

Data preprocessing is very important to transform the raw textual data for analysis. The following steps were implemented:

- **Tokenization:** The reviews were segmented into smaller parts in most cases at the word level to make it easier to analyze.

  For example, The main adjectiveic sentiments are: “This product is amazing!” is being processed and transformed into ['This’, ‘product’, ‘is’, ‘amazing].

- **Stopword Removal:** These content words included simple words such as “the,” “is,” “and” which are not sufficient to determine positive or negative sentiment and NLTK has some prepared list of stopwords.

- **Lemmatization:** Common bases of the words were used for comparison to reduce varied forms of the words to their standard meanings.

  For example, “running” is referred to as “run” and “better” is referred to as “good”.

- **Punctuation and Special Characters:** All the non-significant characters (such as !, #, @ and others) were omitted lest such elements of the text could be of importance.

- **Normalization:** All words were transferred to lower case and phrases such as “don’t” were written as “do not”.

- **Handling Missing Data:** These included reviews with truncated text and were excluded to enhance data quality and ad coherency to the study.

3.2 Feature Engineering

Feature engineering preprocesses textual data converting them to numerical vectors that may be used by machine learning models. The following techniques were applied:

1. **Bag-of-Words (BoW):**
   Converts a text into the feature vector of the word counts.
   Records whether or not certain words are present within a text but does not factor in background.

2. **TF-IDF (Term Frequency-Inverse Document Frequency):**
   Its main idea is assigning weights to words, which can be done with frequencies of words in the document and frequencies of words in the whole collection.
   Can be used in order to draw attention to special keywords and avoid overloading the importance to frequent words.

3. **N-Grams:**
   Records sequences of words (for example bigrams/trigrams to keep the context).
   Example: I retrieved the combination of the terms “great product” and “highly recommend” as the bigrams.

4. **Dimensionality Reduction:**
   Practical approaches to reduce feature space and mitigate high dimensionality issues (e.g., by restricting the list of words to be with no more than 5,000) were also applied.

3.3 Machine Learning Models

3.3.1 Logistic Regression

Description:

Logistic Regression is one of the simplest yet most effective linear models that are used for binary case and also for multiclass case. They are used in estimating the likelihood or risk of an input data to fall under a certain class using logistic function. Therefore, logistic regression is adopted commonly as the base models because of its easy interpretability and simplicity.

Strengths:

- It is easy to implement and computationally optimal robust.
- It works efficiently when the variables and the target labels are related almost linearly.
- A nice property of this model compared to other models where probabilities of the classes are not presented is that it offers a probability for each class.

Limitations:

- Uses feature space where linear relationship between the features and the target is assumed most of the time this is not a true assumption.
- May be sensitive to highly skewed distributions or inability to handle non–linear data distribution or other data distortion effects.
- Concerning its effectiveness, it is sensitive to feature magnitudes, and hence, it requires feature scaling.

3.3.2 Random Forest

Description:

Random Forest is an expanded learning model, based on decision trees, which increases the reliability of forecasts and reduces such a disease as overfitting. At the ensemble level, it compiles the predictions from individual trees either by averaging the values, in case of regression or by a voting system in the case of classification. Random Forest improves upon accuracy and examines the many non-linear associations between features and output by building multiple decision trees with partitions of the data.

Strengths:

- It performs well in dealing with the interaction and quadratic terms among the features.
- Inherent in boosting, AdaBoost is relatively resistant to overfitting, especially when a sufficient number of trees are used (n_estimators).
- Selects features automatically as an important aspect of the training process that aims at outlining important aspects of the data set in question.
- Comparable in its performance for both categorical and continuous variables.

Limitations:

- It is computationally expensive particularly if data is huge or several trees are used.
- Slightly more difficult to explain compared to other models such as the Logistic Regression model.

- Suffers from certain problems like selection of parameters which includes n_estimators, max_depth, min_samples_split, etc.
- May indeed perform poorly for high-dimensional data especially when dominated by irrelevant features.

3.4 Evaluation Metrics:

To assess the performance of the machine learning models, the following metrics were used:

1. Accuracy: Measures the global accuracy of the predictions of this model.
2. Precision: Determines the number of confirmed cases that were predicted to increase.
3. Recall: Measures the percentage of actual positives caught by the model in the whole population of positive samples.
4. F1-Score: The F-score or the harmonic mean of precision and recall, taking care of the trade-off between the two.
5. Confusion Matrices: Computed for each model to show the distribution of the positive , Negative and Neutral Sentiment.

These matrices proved useful in identifying problem areas pertinent to models including misclassification.

3.5 Graphical Representation of the Framework

A flowchart depicting the overall framework design was created to provide a clear visual representation. Below is a description of the process outlined in the flowchart:

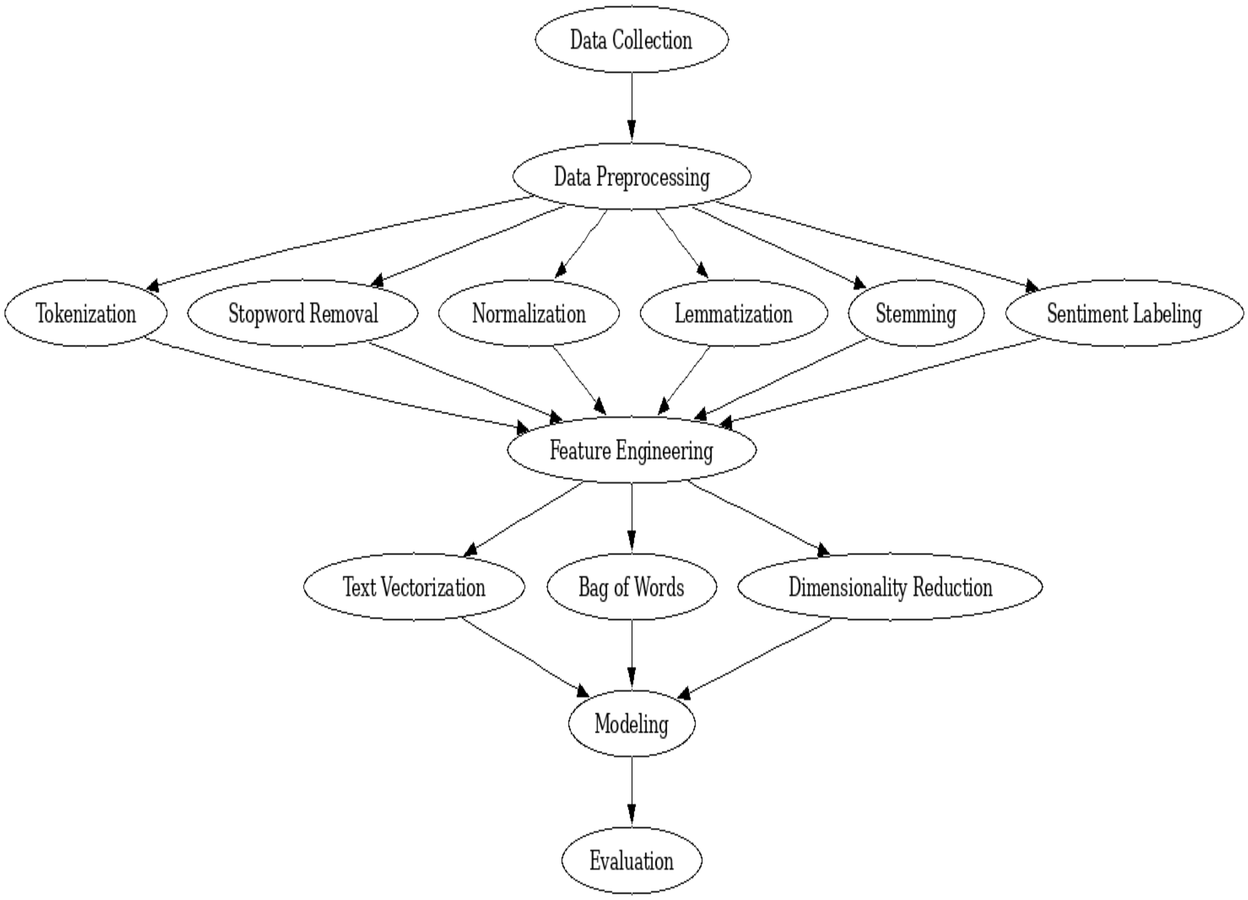


Figure 1: Methodology

1. **Input Data**: Raw textual data is fed into the system.
2. **Preprocessing:** Data is cleaned, tokenized, and transformed into numerical features.
3. **Feature Extraction:** Techniques like TF-IDF and n-grams are applied.
4. **Model Training:** Preprocessed data is split into training and test sets, and models are trained using machine learning algorithms.
5. **Evaluation:** Model predictions are compared against ground truth labels using evaluation metrics.

3.6 Visualization Tools

1. Matplotlib

- ❖ Purpose: A versatile library used for creating static, animated, and interactive visualizations in Python.
- ❖ Usage:
    - ➢ To generate standard plots such as bar charts, line charts, and scatterplots.

  - ➢ Example: A bar chart was created to visualize the distribution of sentiments in the dataset, showing the counts of positive, negative, and neutral sentiments.

2. Seaborn

- ❖ Purpose: A Python data visualization library based on Matplotlib, offering a high-level interface for creating attractive and informative statistical graphics.
- ❖ Usage:
  - ➢ To create aesthetically pleasing bar plots, heatmaps, and pair plots.
  - ➢ Example: A bar plot was used to represent the sentiment distribution across categories, leveraging Seaborn's style and color palette to enhance interpretability.

3. WordCloud

- ❖ Purpose: A visualization tool that generates a cloud of words based on their frequency in a dataset.
- ❖ Usage:
  - ➢ To visually explore the most frequent terms in the dataset for each sentiment category (positive, negative, and neutral).
  - ➢ Example: WordClouds were generated to highlight keywords in customer reviews, with frequent words appearing larger in size.

4. Confusion Matrix

- ❖ Purpose: A matrix used to evaluate the performance of classification models by showing the number of correct and incorrect predictions for each class.
- ❖ Usage:
  - ➢ Created using Matplotlib to analyze the predictions of machine learning models (e.g., Logistic Regression and Random Forest).
  - ➢ Example: Visualized misclassifications for better understanding of model strengths and weaknesses.

5. Sentiment Distribution

- ❖ Purpose: Exploratory data visualization to understand the distribution of sentiment labels in the dataset.
- ❖ Usage:
  - ➢ Bar charts were plotted to show the counts of positive, negative, and neutral reviews.
  - ➢ This helped provide an overview of customer sentiment trends in the dataset.

4.0 Evaluation and Analysis

Evaluation phase included low of the performance of the machine learning models in when using text as the parameter, the sentiments in the samples should be classified as well as the trends that is stored in the dataset. By combining metrics and analyses in the form of these visualizations, this section ensures that the reader gets a deeper insight of the model s characteristics and sentiment. structure, types, and language usage in the political communication.

4.1 Model Performance

The performance of the machine learning models was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, confusion matrices were generated to visualize classification performance across the sentiment categories (positive, negative, and neutral).

4.1.1 Logistic Regression

Performance Overview:
The Logistic Regression model provided a baseline for sentiment classification. Despite its simplicity, the model performed effectively, particularly for neutral sentiments, which constituted the majority of the dataset.

Key Metrics:

- **Accuracy**: Achieved an accuracy of [0.9108409321175278], reflecting balanced performance across sentiment categories.
- **Precision**: [0.90] for positive, negative, and neutral categories.
- **Recall**: [0.91], with the model excelling in detecting neutral sentiments.
- **F1-Score**: [0.91], showcasing a balance between precision and recall.

Confusion Matrix: A confusion matrix was generated to highlight the model's performance across sentiment categories, revealing its strengths in handling neutral sentiments and weaknesses in misclassifying subtle linguistic overlaps.

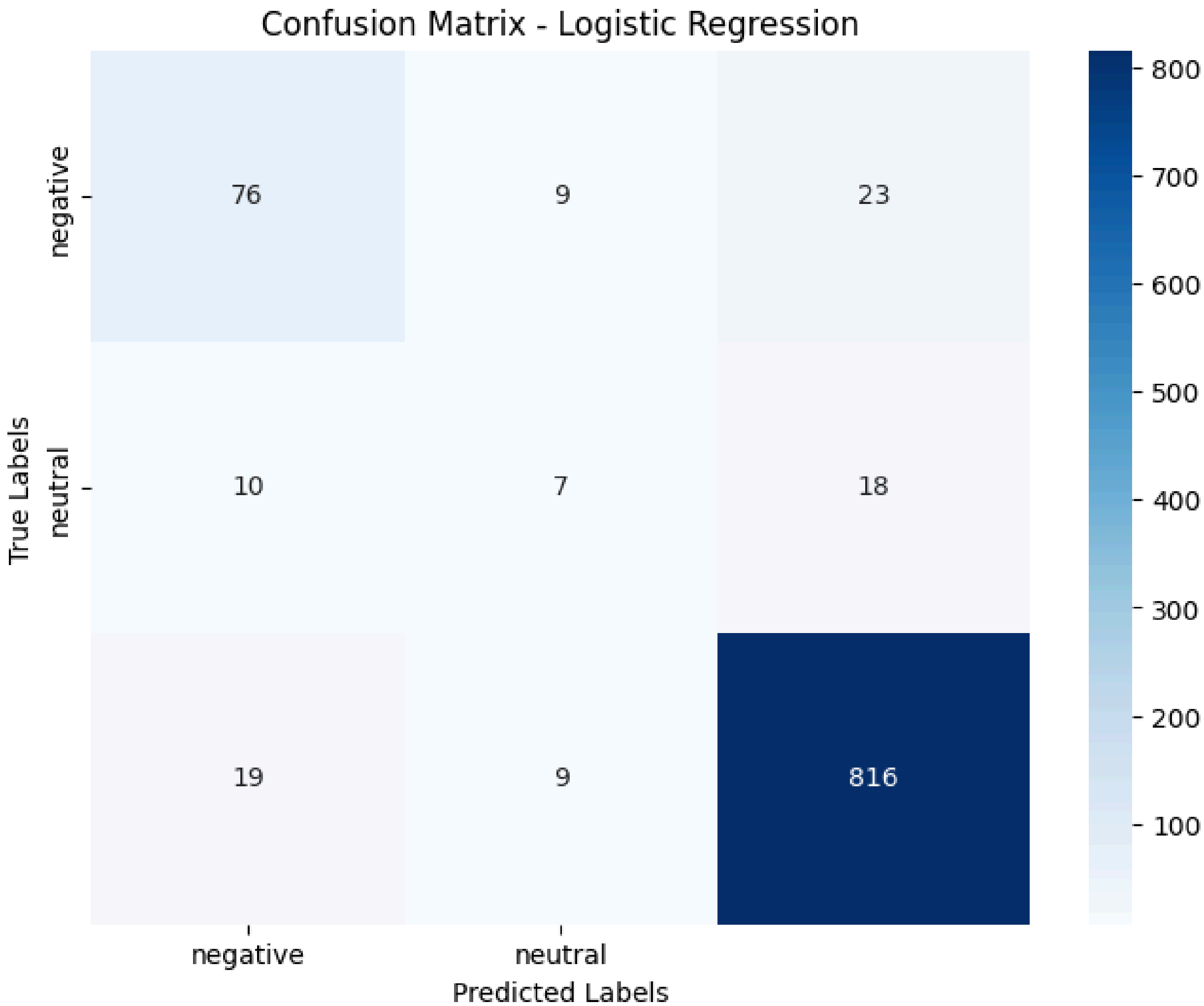


Figure 2: confusion matrix for Logistic Regression

4.1.2 Random Forest

Performance Overview:
The Random Forest model demonstrated robust performance by capturing non-linear relationships in the data. However, its performance was slightly hindered by misclassifications in minority classes (e.g., negative sentiments).

Key Metrics:

- **Accuracy**: Achieved an accuracy of [0.9159067882472138], slightly higher than Logistic Regression.
- **Precision**: [0.89], particularly high for neutral sentiments.
- **Recall**: [0.92], with challenges in identifying negative sentiments.
- **F1-Score**: [0.90], indicating improved performance over Logistic Regression for positive and neutral classes.

Confusion Matrix: The confusion matrix revealed that while Random Forest excelled in classifying neutral and positive sentiments, it occasionally misclassified negative sentiments as neutral or positive. This highlights the need for further tuning or additional techniques to address class imbalance.

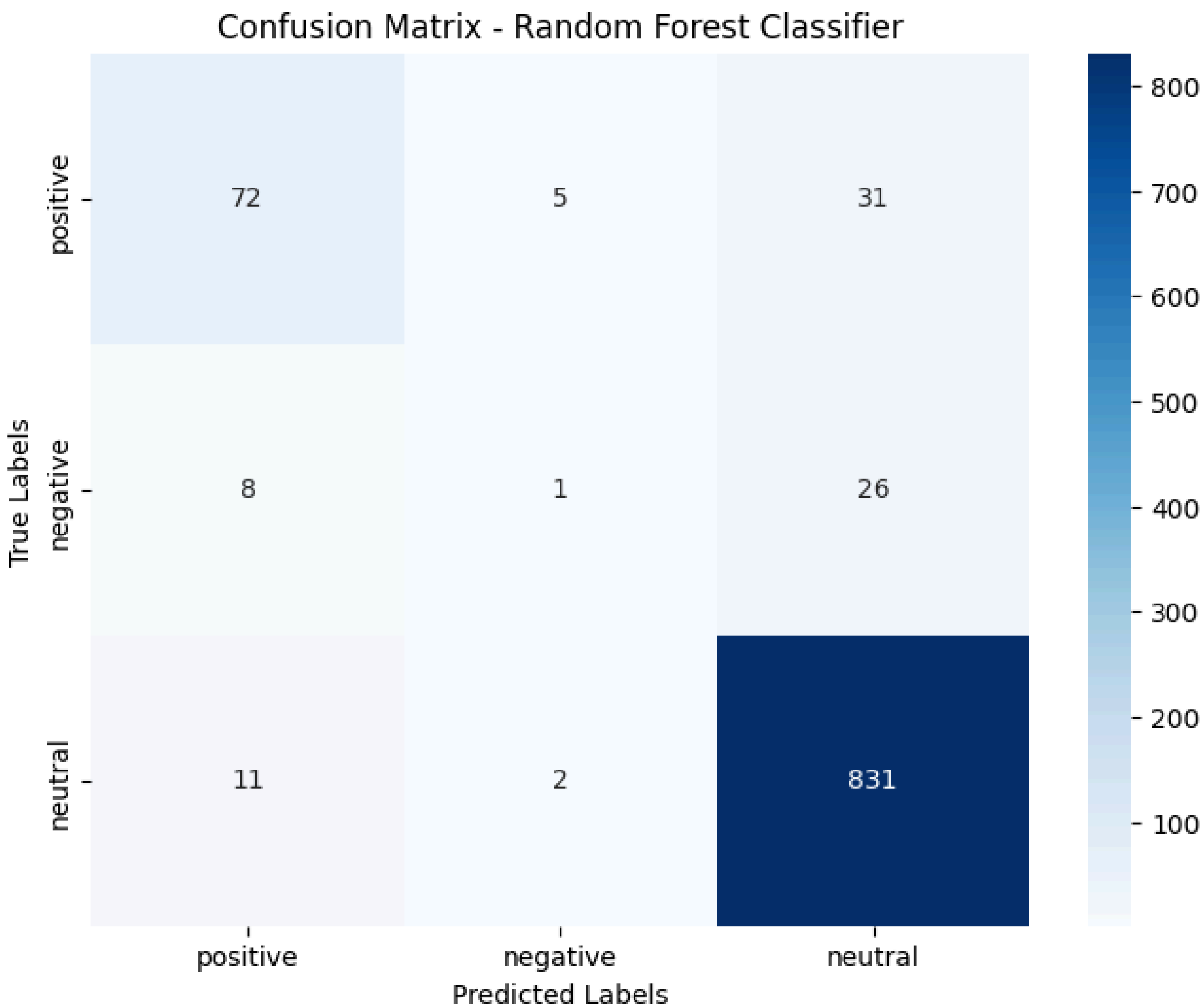


Figure 3: confusion matrix for Random Forest

4.2 Comparative Performance Analysis :

The comparison between Logistic Regression and Random Forest models highlights their respective strengths and limitations in sentiment classification. Both models were evaluated using key performance metrics: Accuracy, Precision, Recall, and F1-Score.

Performance Metrics:

Metric	Logistic Regression	Random Forest
Accuracy	0.91	0.92
Precision	0.90	0.89
Recall	0.91	0.92
F1-Score	0.91	0.90

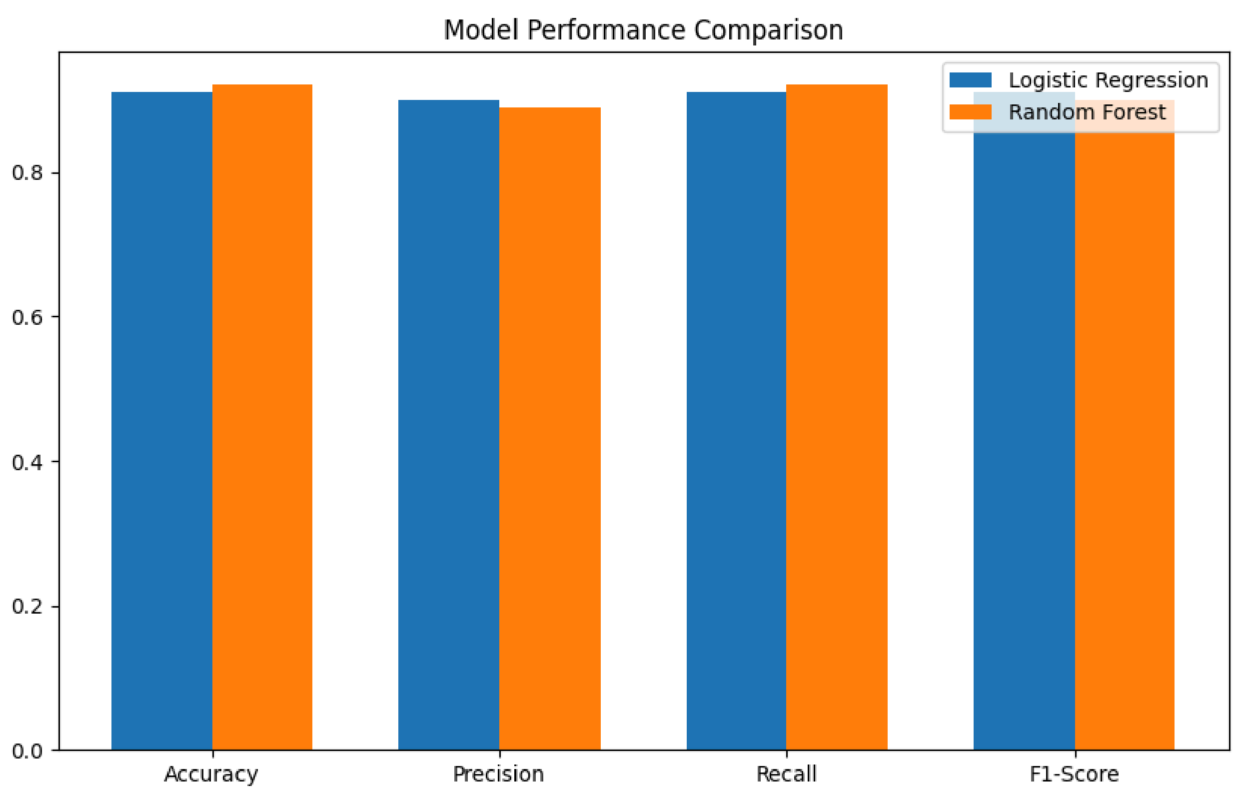


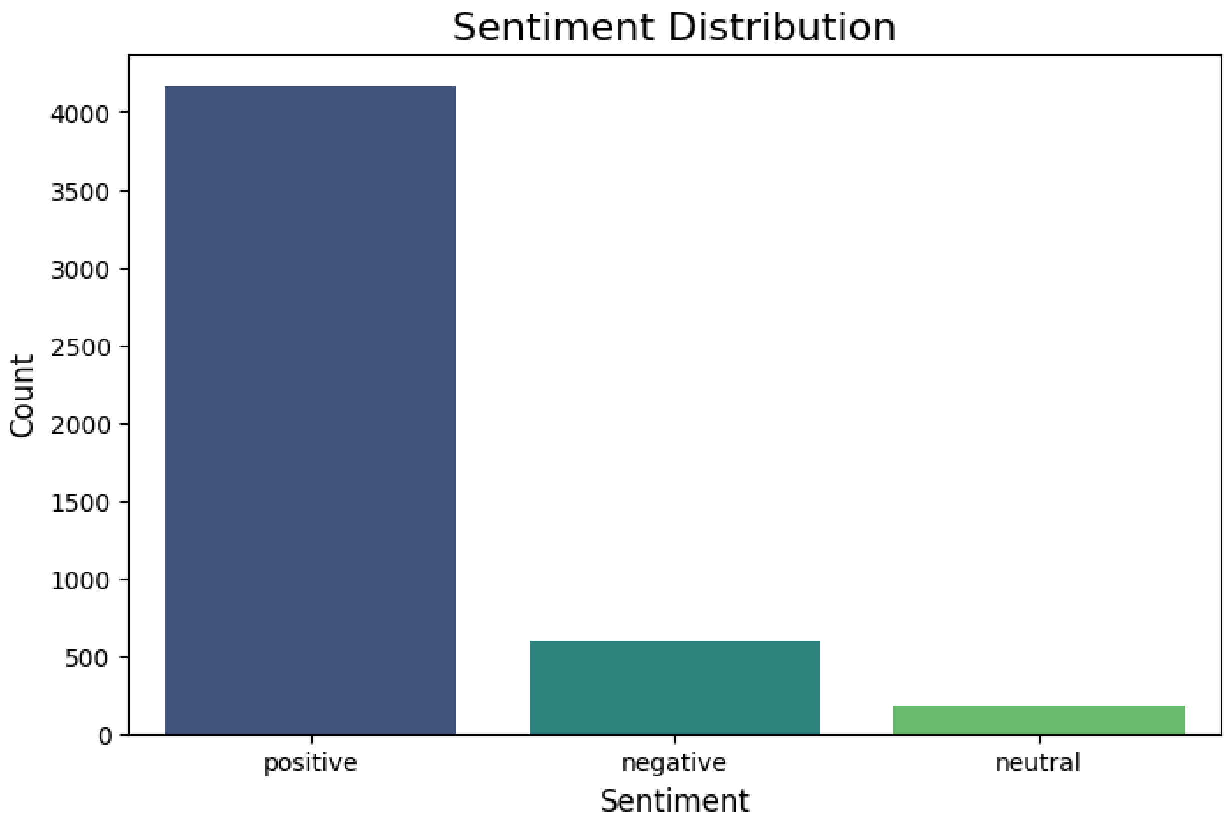
Figure 4: Model Performance Comparison

Insights:

1. **Overall Accuracy**:
   - Random Forest achieved slightly higher accuracy (92%) compared to Logistic Regression (91%), indicating its robustness in capturing non-linear patterns in data.
2. **Precision**:
   - Logistic Regression demonstrated better precision (90%) for classifying individual sentiments compared to Random Forest (89%), which suggests a lower rate of false positives.
3. **Recall**:
   - Random Forest excelled in recall (92%), indicating its ability to detect more relevant instances, particularly in minority classes like negative sentiments.
4. **F1-Score**:
   - Logistic Regression had a marginally better F1-Score (91%), reflecting its balanced performance across precision and recall.

4.3 Sentiment Distribution

Analyzing the sentiment distribution of the Flipkart reviews provides insights into the overall customer feedback and highlights areas of satisfaction or dissatisfaction.



Distribution Overview

The dataset contains reviews classified into three sentiment categories: positive, negative, and neutral. The counts for each category are as follows:

- **Positive Sentiments**: 4,164 reviews
- **Negative Sentiments**: 595 reviews
- **Neutral Sentiments**: 174 reviews

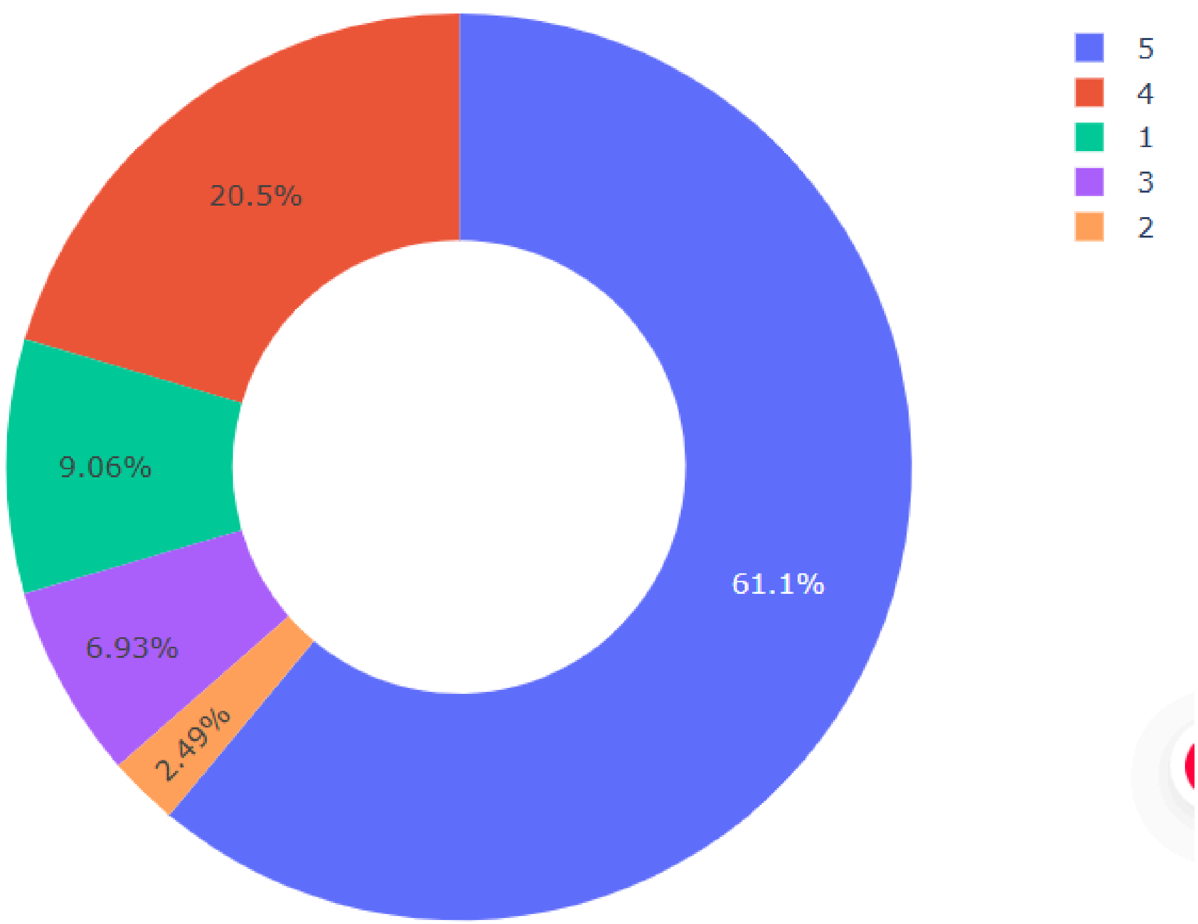
Key Observations

1. **Positive Sentiments**:
    - The majority of the reviews (4,164) fall into the positive sentiment category, accounting for approximately **85.6%** of the dataset. This indicates that most customers are satisfied with their purchases and experiences.
2. **Negative Sentiments**:
    - A smaller portion of the reviews (595), approximately **12.2%**, reflect dissatisfaction. This highlights areas where customers encountered issues or unmet expectations.

3. **Neutral Sentiments**:
    - Neutral reviews (174) constitute **2.2%** of the dataset, representing feedback that lacks strong emotional tones, often descriptive or factual.

4.4 Rating Overview

The customer rating distribution provides critical insights into the overall satisfaction levels of customers. Ratings are divided into five categories, ranging from 1 (very dissatisfied) to 5 (very satisfied). The distribution is visualized in the donut chart above, where each section represents the proportion of reviews with a specific rating.

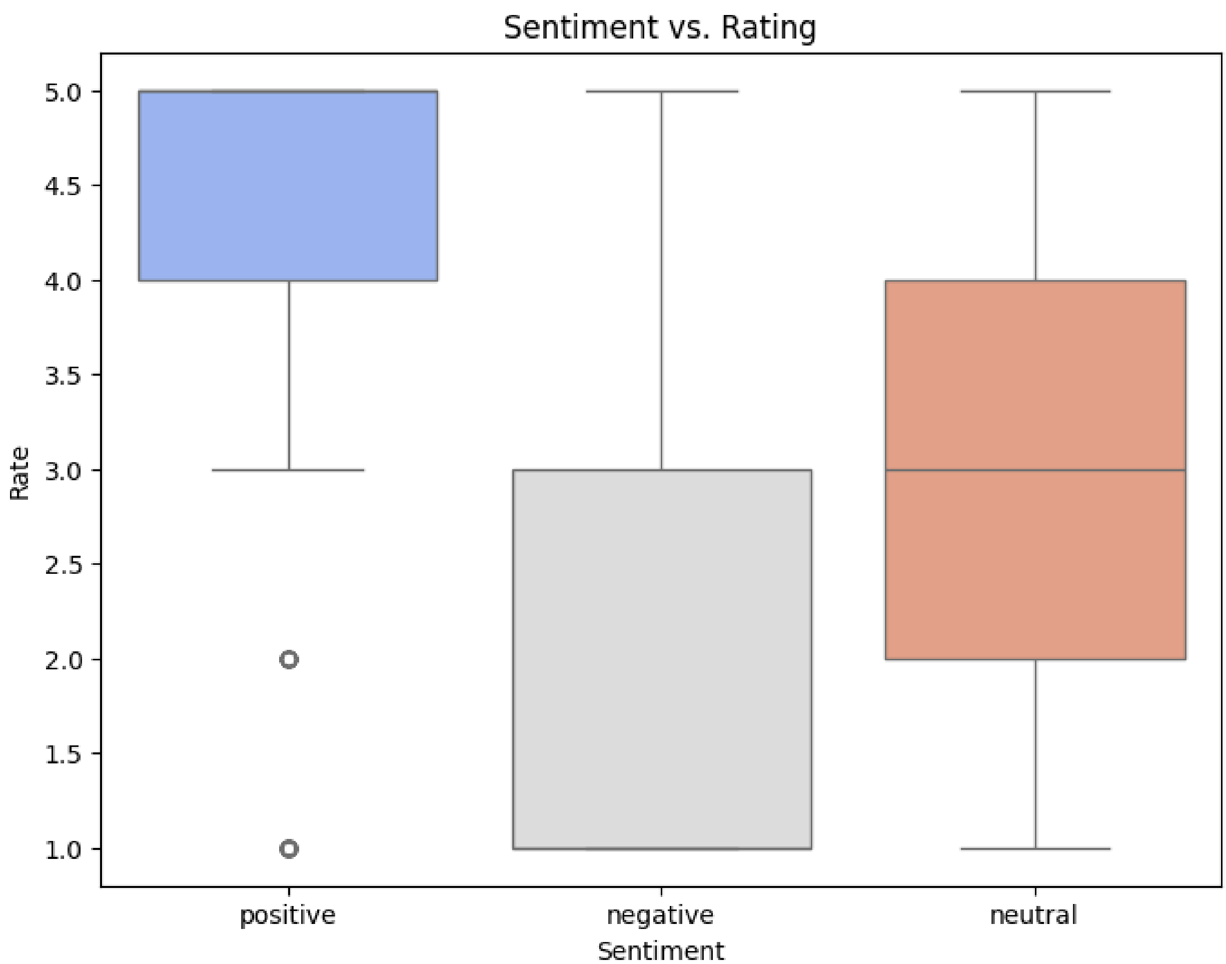


Rating Distribution:

- **5-Star Ratings**: **61.1%**
  - The majority of customers rated their experience as excellent, highlighting high satisfaction with the products or services.
- **4-Star Ratings**: **20.5%**
  - A significant portion of customers gave good ratings, suggesting general satisfaction but with minor concerns.
- **3-Star Ratings**: **6.93%**
  - Neutral ratings indicate average experiences, where customers might not have been strongly satisfied or dissatisfied.
- **2-Star Ratings**: **2.49%**
  - These ratings reflect dissatisfaction, likely caused by unmet expectations or minor issues.
- **1-Star Ratings**: **9.06%**
  - The lowest ratings indicate severe dissatisfaction, pointing to significant issues with products or services.

4.4.1 Sentiment vs Rating Comparisons :

The comparison between sentiment categories (positive, negative, neutral) and their associated ratings provides valuable insights into how customer sentiments align with their given ratings. The boxplot above illustrates the distribution of ratings for each sentiment category



1. **Alignment Between Sentiment and Rating**:
    - Positive sentiments strongly align with higher ratings, reflecting customer satisfaction.
    - Negative sentiments correlate with lower ratings, capturing dissatisfaction effectively.
    - Neutral sentiments show diverse ratings, as expected, given the lack of strong emotional tones in neutral feedback.
2. **Outliers**:
    - Outliers in the positive and negative sentiment categories indicate occasional discrepancies between sentiment polarity and the assigned rating. These outliers could arise from factors such as:
        - Inconsistent reviewer behavior.
        - Miscommunication in text sentiment versus rating.
3. **Opportunities for Improvement**:

- Understanding cases where positive sentiments correspond to lower ratings can help businesses address mismatched expectations or product quality concerns.
- Addressing variability in negative and neutral sentiments can enhance customer experience and foster better alignment between sentiments and ratings.

4.5 WordCloud Analysis

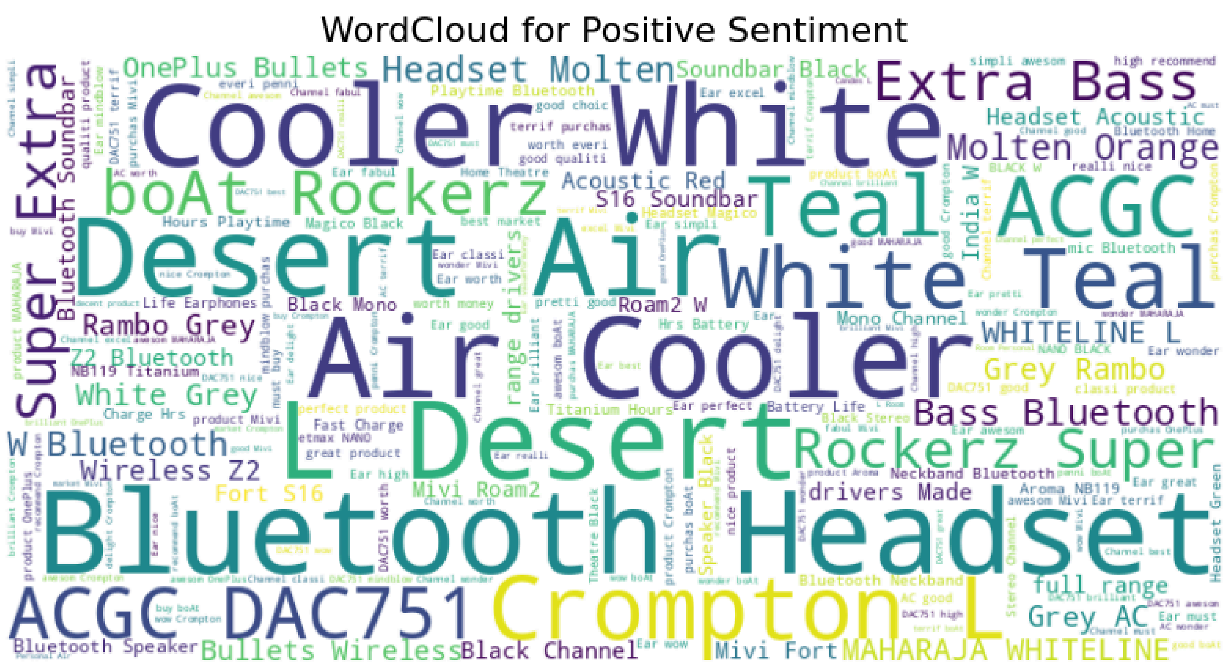


Figure 8 : Wordcount positive Sentiment

The WordCloud for **positive sentiments** highlights frequently mentioned terms, emphasizing customer satisfaction with products like **Bluetooth Headsets**, **Air Coolers**, and brands such as **Crompton** and **boAt**. Features like **Extra Bass** and aesthetic elements like **White** and **Teal** are prominently appreciated. This visualization provides actionable insights into customer preferences, helping businesses focus on enhancing popular features and promoting well-received products to sustain high satisfaction levels.

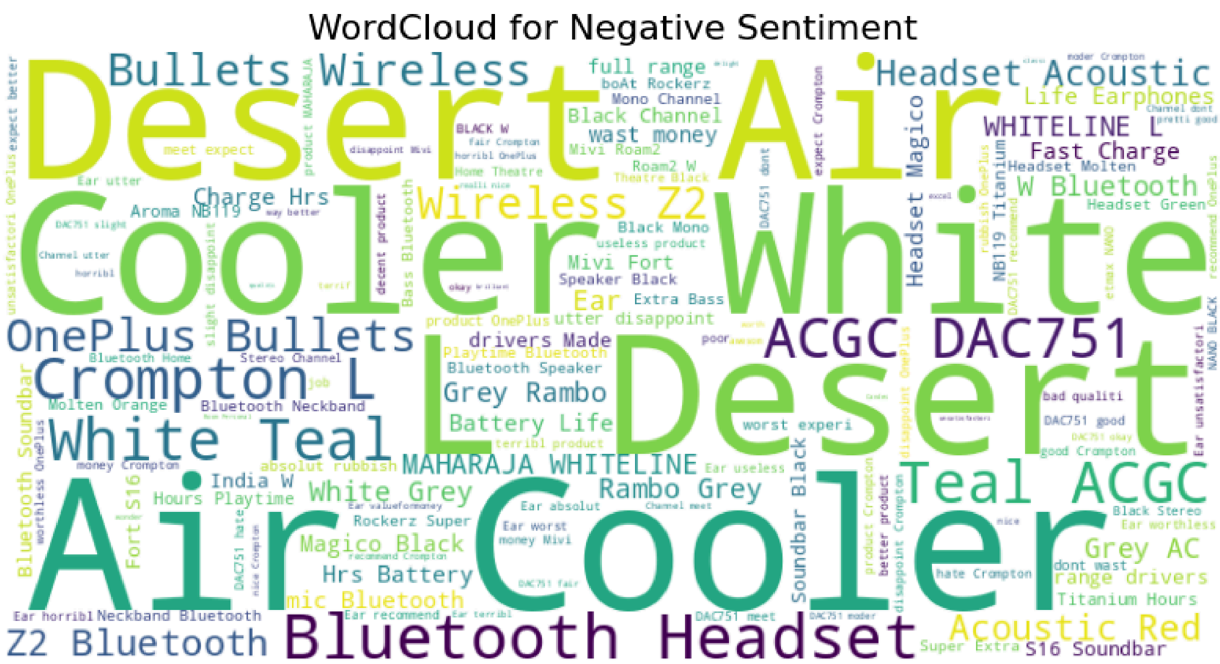


Figure 9 : Wordcount negative Sentiment

The WordCloud for **negative sentiments** highlights frequently mentioned terms associated with customer dissatisfaction. Products like **Air Coolers**, **Bluetooth Headsets**, and brands such as **Crompton** and **OnePlus** are prominently featured. Negative feedback often focuses on issues like **battery life**, **product quality**, and **delivery expectations**. This visualization provides insights into areas needing improvement, enabling businesses to address customer concerns and enhance overall satisfaction.

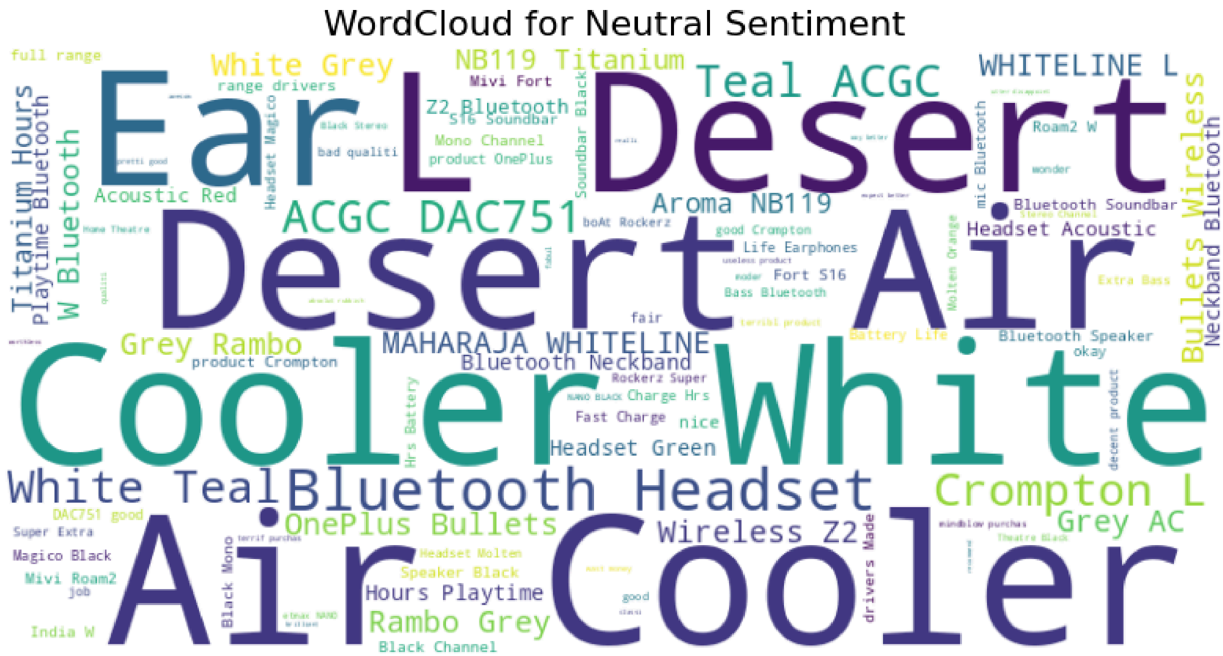


Figure 10 : Wordcount neutral Sentiment

The WordCloud for **neutral sentiments** highlights frequently mentioned terms such as **Air Coolers**, **Bluetooth Headsets**, and **Desert Air**, alongside brands like **Crompton** and **OnePlus**. Neutral feedback typically reflects descriptive or factual observations without strong emotional tones, focusing on features like **battery life**, **color options** (e.g., **White** and **Teal**), and **product usage**. This visualization provides an overview of commonly discussed products and features, helping businesses refine customer communication and address any ambiguities in product descriptions.

5.0 Conclusion

The sentiment analysis of Flipkart customer reviews provided valuable insights into customer satisfaction, preferences, and areas for improvement. Using Natural Language Processing (NLP) and machine learning techniques, the analysis categorized customer sentiments into positive, negative, and neutral categories, revealing that positive sentiments overwhelmingly dominated the dataset, reflecting high levels of satisfaction.

The analysis highlighted key trends, such as the popularity of products like Bluetooth Headsets and Air Coolers and the appreciation for features like sound quality and design aesthetics. Negative sentiments, while fewer, identified areas such as product quality and delivery concerns that require attention. Neutral sentiments offered descriptive feedback, providing additional context for understanding customer experiences.

Comparing machine learning models, Random Forest demonstrated slightly better performance in capturing complex patterns in the data, while Logistic Regression served as a reliable baseline with competitive results. Visual tools such as WordClouds, bar charts, and confusion matrices effectively illustrated sentiment trends and model performance.

This project underscores the importance of sentiment analysis for e-commerce platforms in understanding customer feedback and guiding data-driven decisions. By addressing the concerns highlighted in negative reviews and leveraging insights from positive sentiments, businesses can enhance customer satisfaction and loyalty, ultimately driving growth and competitive advantage.

6.0 Appendix

Dataset

https://www.kaggle.com/code/amirmotefaker/flipkart-reviews-sentiment-analysis

PLAGIARISM STATEMENT

I certify that this assignment/report is my own work, based on my personal study and/or research and that I have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication. I also certify that this assignment/report has not previously been submitted for assessment in any other subject, except where specific permission has been granted from all unit coordinators involved, or at any other time in this unit, and that I have not copied in part or whole or otherwise plagiarized the work of other students and/or persons.

Name : Mutasim Billah

Date :.8-01-2025

Signature :Mutasim