Name

- Afrose Ahamed Fathima

Course

- Advanced Certification in Data Science and Artificial Intelligence

Capstone Project

Reviews

Date: 08-10-2024

Table of Contents

Acknowledgment

I would like to express my sincere gratitude to the Intellipaat Team and IIT Madras for providing me with the opportunity to enhance my skills and knowledge through their comprehensive course. The curriculum, coupled with expert guidance, has significantly contributed to my understanding of data science, machine learning, and artificial intelligence. I appreciate the dedication and support of the instructors and the valuable resources that were made available throughout the program. I am eager to apply what I have learned in future endeavours.

Problem Statement

You are working in an e-commerce company, and your company has put forward a task to analyze the customer reviews for various products. You are supposed to create a report that classifies the products based on the customer reviews.

1. Find various trends and patterns in the reviews data, create useful insights that best describe the product quality.
2. Classify each review based on the sentiment associated with the same.

Problem Objective

The objective of this project is to analyze customer reviews from an e-commerce platform and classify the products based on the sentiment associated with those reviews. By building a machine learning model and conducting sentiment analysis, the goal is to identify trends and patterns in customer feedback, understand product quality, and provide useful insights that describe how different products are perceived by customers.

Data Description

Feature Name	Description	Data Type
Id	Record ID	Integer
ProductId	Product ID	String
UserId	User ID who posted the review	String
ProfileName	Profile name of the User	String
HelpfullnessNumerator	Numerator of the helpfulness of the review	Integer
HelpfullnessDenominator	Denominator of the helpfulness of the review	Integer
Score	Product Rating	Integer
Time	Review time in timestamp	Integer
Summary	Summary of the review	String
Text	Actual text of the review	String

The data contains 568454 rows and 10 columns out of which 54 are missing values.

Data Pre-processing Steps and Inspiration

1. **Handling Missing Values:**
   - ProfileName: There are 26 missing values in the dataset.
   - Summary: There are 27 missing values in the dataset.

   The missing values are removed from the dataset.

2. **Outlier Detection**:
   - Techniques like boxplots are used to detect and manage outliers effectively. Outliers in HelpfulnessNumerator, HelpfulnessDenominator, Score, and Time are treated by removing them.

3. **Text Cleaning**: The following text cleaning techniques are used:
    - **Tokenization**: Split text into individual words.
    - **Stopword Removal**: Removed common words (like "the", "and") that do not contribute to meaning.
    - **Lowercasing**: Converted all text to lowercase for consistency.
    - **Punctuation Removal**: Removed punctuation and special characters from the review text.
    - **Lemmatization**: Converted words to their base form (e.g., "running" to "run").

4. **Text Vectorization:**
    - Used TF-IDF (Term Frequency-Inverse Document Frequency) to transform the cleaned text into numerical features that can be fed into machine learning models.

5. **Feature Engineering**:
    - Created new features such as review length (word count) and sentiment polarity using the TextBlob library.

6. **Target Variable Creation**:
    - Converted product ratings into binary classification labels:
        1. **High Quality** for ratings 4 and above.
        2. **Low Quality** for ratings below 4.

Inspiration:

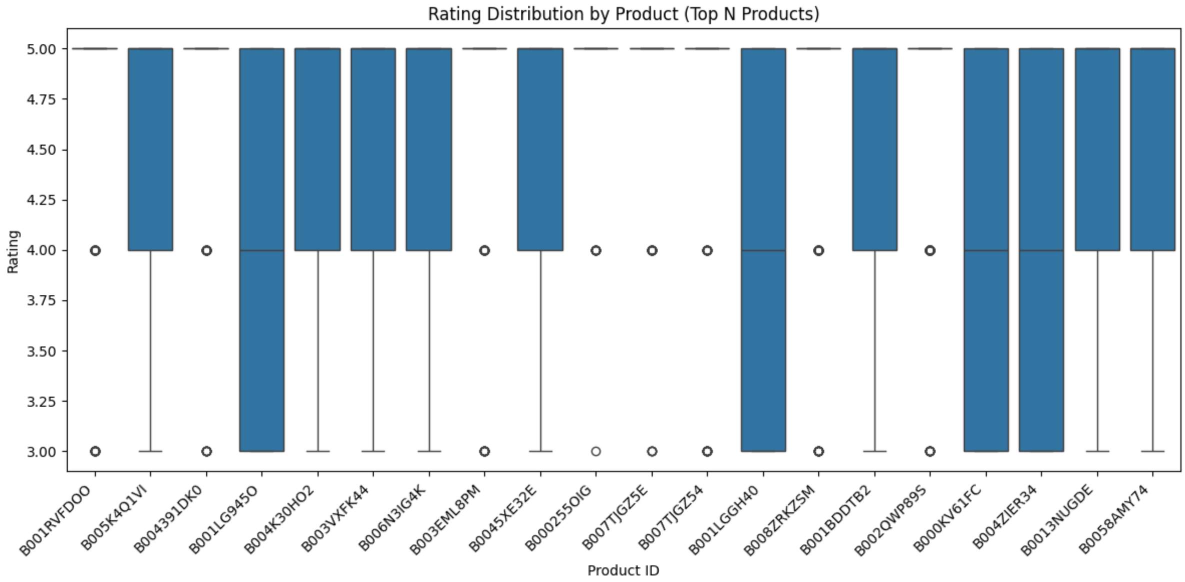
Sentiment Analysis: Leveraged sentiment polarity to classify reviews based on positive, negative, or neutral feedback.

Feature Importance: Used machine learning (Random Forest) to find important words and patterns that impact product quality.

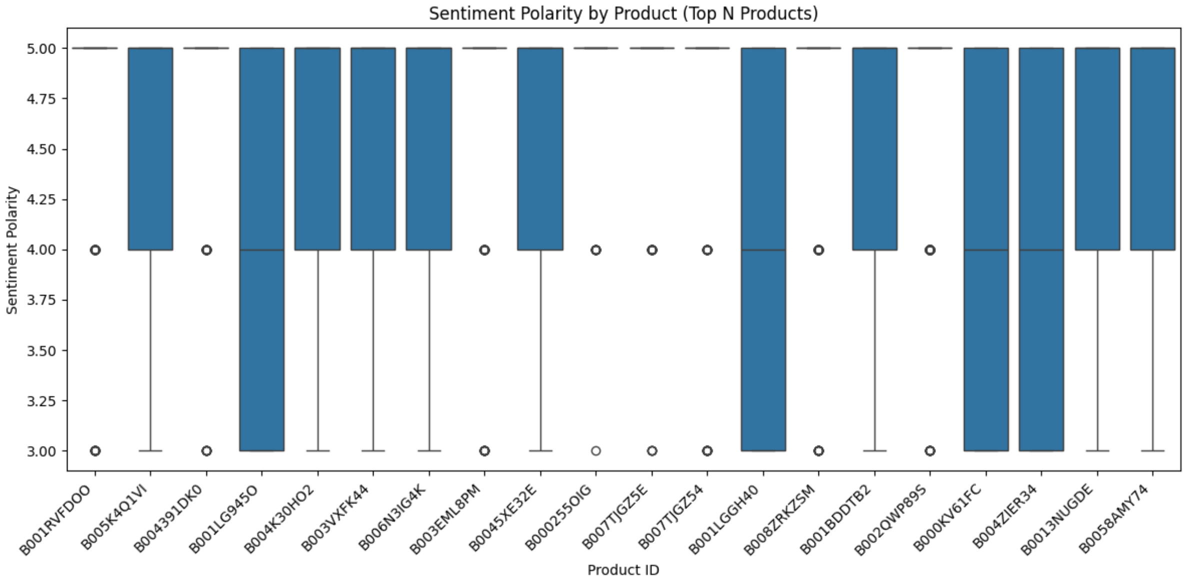
Trend Analysis: Explored relationships between review sentiment, ratings, and other key features to gain insights into product quality.

Data Visualization

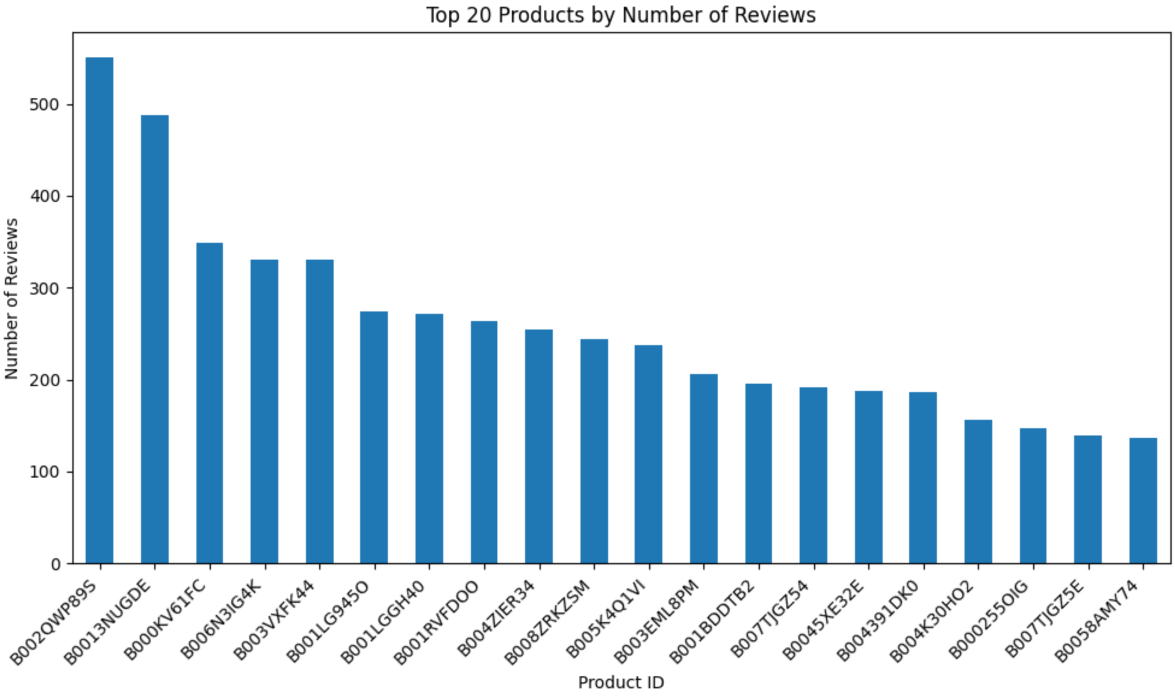
Rating Distribution by top products:



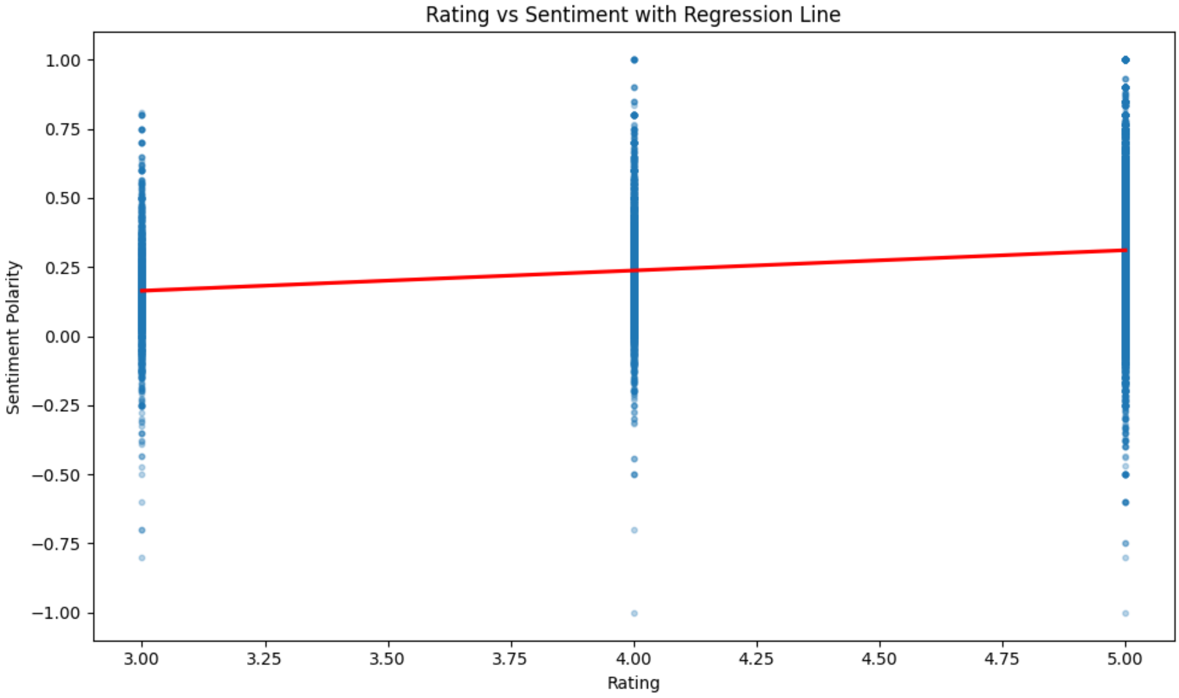
Sentiment Polarity by Top Products:



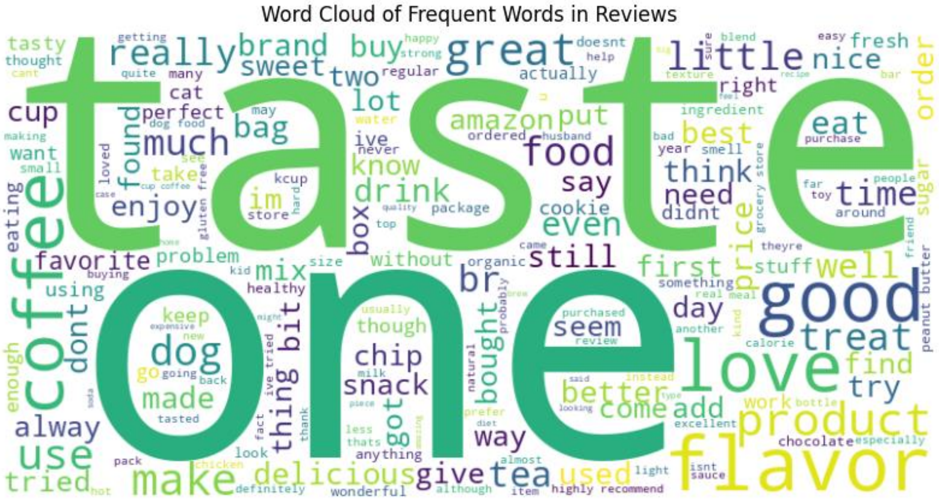
Number of Reviews by Top Products:



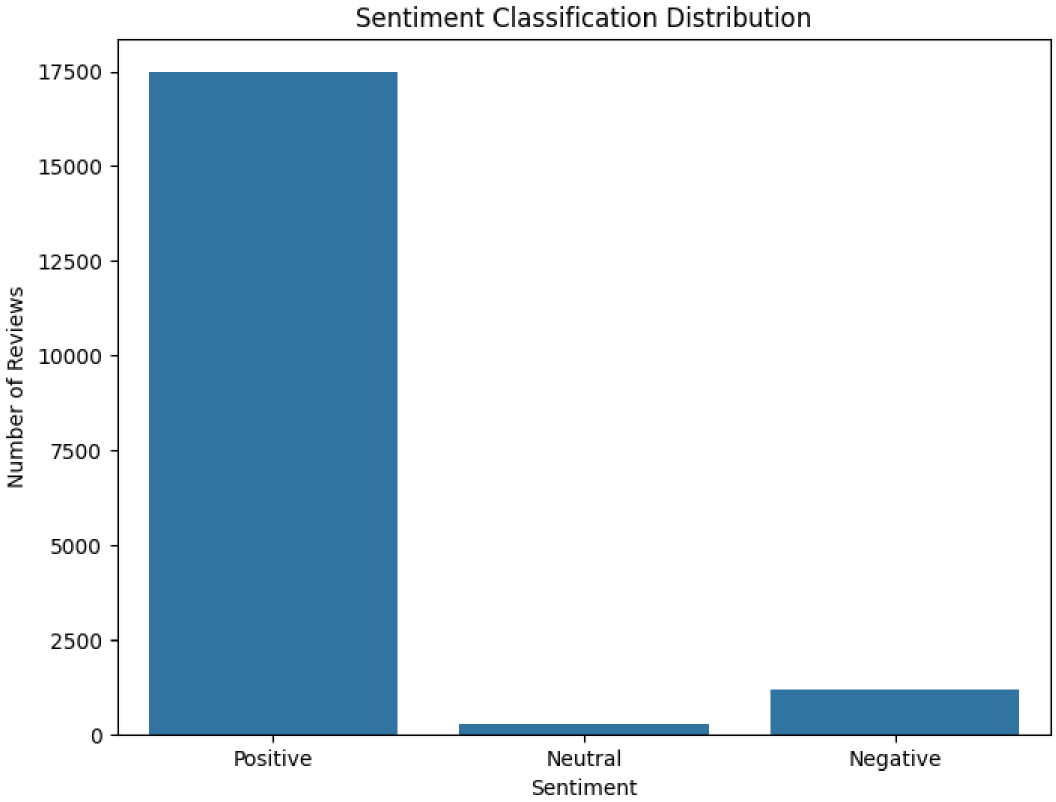
Rating vs Sentiment:



Word Cloud of Frequent Words in Reviews



Sentiment Classification Distribution:



Choosing the Algorithm for the Project

For this project, **Random Forest** was selected as the primary algorithm to classify products based on customer reviews and predict product quality.

Motivation and Reasons For Choosing the Algorithm

The motivation for choosing **Random Forest** as the algorithm for this project is based on the following reasons:

1. **Accuracy and Performance**:
   - Random Forest is known for delivering high accuracy by combining multiple decision trees and reducing overfitting, which is crucial when dealing with complex, text-based review data.
2. **Handling High-Dimensional Data**:
   - With text data vectorized into many features (using TF-IDF), Random Forest can handle high-dimensional datasets effectively without the need for heavy feature engineering or dimensionality reduction.
3. **Feature Importance**:
   - A key motivation was its ability to rank feature importance, which helps identify the most influential words or aspects of reviews driving product quality predictions. This adds transparency and interpretability to the model.
4. **Versatility**:
   - Random Forest can handle both categorical and continuous data, making it flexible for combining textual features (reviews) with numerical features (ratings).

Assumptions

1. **Quality of Reviews**:

   It is assumed that customer reviews provide reliable and honest feedback regarding the products. The sentiment expressed in the reviews is reflective of the actual product quality.

2. **Textual Representation**:

   The assumption is made that the cleaned and processed review text effectively captures the sentiments and opinions of customers. Proper text preprocessing (e.g., removing stopwords, and punctuation) ensures meaningful information is retained.

3. **Independence of Reviews**:

   It is assumed that each review is independent of others. The analysis treats each review as a standalone input, which may not account for the influence of prior reviews on a customer's sentiment.

4. **Feature Relevance**:

   It is assumed that the features derived from the reviews (e.g., sentiment scores, review lengths) are relevant and meaningful for predicting product quality.

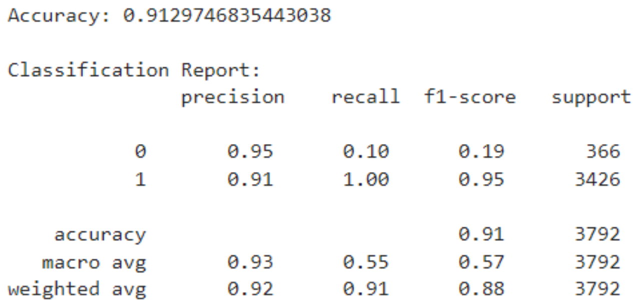
Model Evaluation and Techniques

1. **Train-Test Split**:

The dataset was divided into training and testing sets, typically using an 80/20 split. This allows for model training on one subset of data and evaluation on an unseen subset, ensuring that the model's performance is not overestimated.

2. **Performance Metrics**:
   a. **Accuracy**: The proportion of correctly predicted instances out of the total instances.
   b. **Precision**: The proportion of true positive predictions relative to the total positive predictions made (i.e., how many predicted positives were actually positive).
   c. **Recall (Sensitivity)**: The proportion of true positive predictions relative to the actual positives in the dataset (i.e., how many actual positives were correctly identified).
   d. **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two metrics, especially useful when the class distribution is imbalanced.
   e. **Confusion Matrix**: A matrix that provides insight into the model's performance by showing true positives, true negatives, false positives, and false negatives. It helps visualize where the model is making mistakes.

Below is the Performance Metrics:

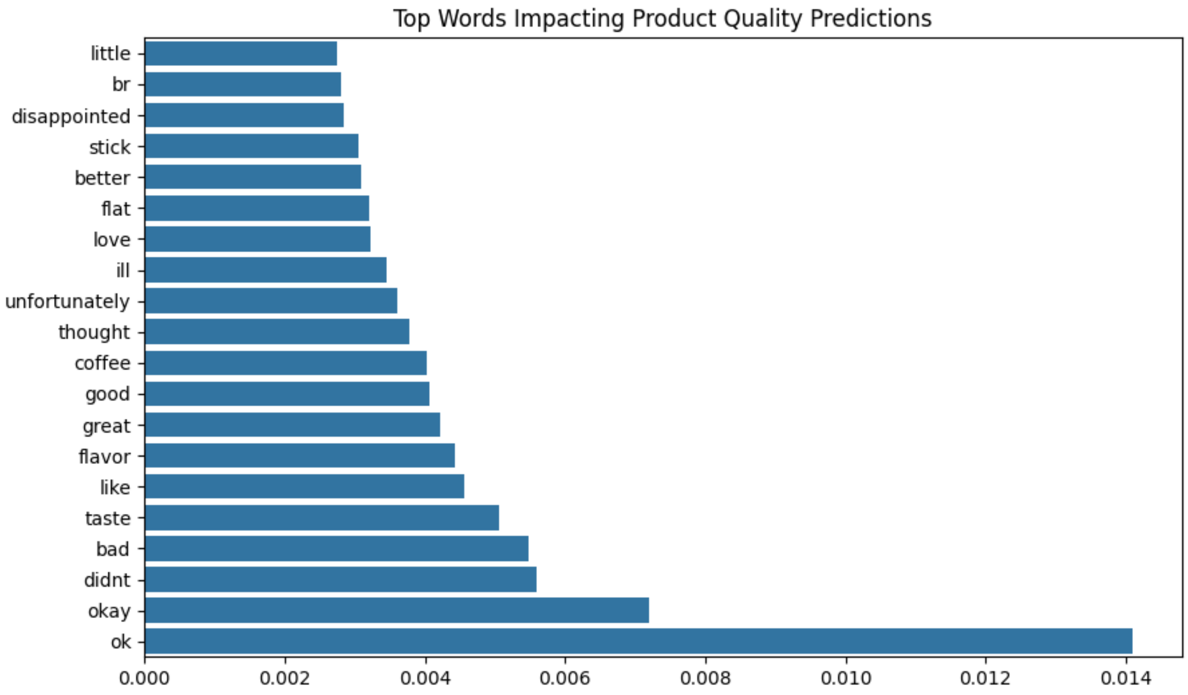


```
Accuracy: 0.9129746835443038

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.10      0.19       366
           1       0.91      1.00      0.95      3426

    accuracy                           0.91      3792
   macro avg       0.93      0.55      0.57      3792
weighted avg       0.92      0.91      0.88      3792
```

An accuracy of 91% indicates that the model correctly classifies 91 out of every 100 customer reviews. This suggests that the Random Forest model performs exceptionally well in distinguishing between different product quality classes (e.g., high quality vs. low quality).

3. **Feature Importance Analysis**:

Random Forest's feature importance scores were analyzed to understand which words or features significantly influenced the model's predictions. This analysis aids in interpreting the model and deriving insights from the reviews.



Inferences from the Same

1. The model's high accuracy indicates a strong confidence level in classifying customer reviews into high-quality and low-quality product categories. This can help stakeholders trust the model's predictions for strategic decision-making.
2. The model's success suggests that the features used (e.g., sentiment scores, review text) are effective in capturing the nuances of customer feedback. This highlights the importance of thoughtful feature engineering and selection in model performance.
3. High accuracy allows for effective segmentation analysis. Businesses can segment their offerings based on product quality and tailor their marketing strategies to different customer demographics or preferences.

Future Possibilities of the Project

1. **Real-Time Sentiment Analysis**:

   Implementing real-time sentiment analysis could allow businesses to monitor customer feedback as it comes in. This would enable quicker responses to negative reviews and help maintain a positive brand image.

2. **Expansion to Multi-Channel Reviews**:

   Extending the analysis to include reviews from multiple channels (social media, forums, and review sites) can provide a more comprehensive understanding of customer sentiment and product performance across different platforms.

3. **Predictive Analytics**:

   Utilizing predictive analytics to forecast customer sentiment trends could help businesses anticipate shifts in consumer behavior and adapt their strategies accordingly, staying ahead of market changes.

4. **Personalization of Recommendations**:

   Leveraging sentiment analysis to personalize product recommendations based on customer feedback could enhance user experience and increase conversion rates.

5. **Integration with Inventory Management**:

   Integrating sentiment analysis with inventory management systems could help businesses align stock levels with customer demand and sentiment, optimizing supply chain efficiency.

Conclusion

The customer review analysis project has successfully harnessed machine learning and natural language processing to classify customer sentiments and assess product quality, achieving high accuracy with the Random Forest model. This analysis has provided actionable insights into customer preferences and pain points, guiding businesses in making informed decisions regarding product enhancements and marketing strategies. By identifying key features that influence perceptions of quality, the project enables organizations to prioritize improvements and maintain a competitive edge in the market.