SENTIMENT ANALYSIS OF RESTAURANT REVIEWS

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

HARSHINI AKSHAYA A S (2116220701088)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"SENTIMENT ANALYSIS OF RESTAURANT REVIEWS"** is the bonafide work of **"HARSHINI AKSHAYA A S (2116220701088)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on ______________________

Internal Examiner **External Examiner**

ABSTRACT

In the realm of sentiment analysis, the development of a robust model is crucial for extracting meaningful insights from the vast array of restaurant reviews available online. The chosen approach of employing the Logistic Regression algorithm, along with essential packages like numpy and pandas, reflects a commitment to a data-driven methodology. This model not only aims to categorize reviews into positive or negative sentiments but also addresses the inherent challenges posed by sarcastic remarks and misspelled words, enhancing its accuracy and reliability. By delving into the nuances of language, the model aspires to provide restaurants with a nuanced understanding of customer feedback, enabling them to make informed decisions to elevate their overall performance.

The significance of this sentiment analysis model extends beyond mere classification. Its implementation holds the potential to revolutionize how restaurants approach customer satisfaction. By leveraging insights gained from positive and negative reviews, establishments can tailor their strategies to amplify positive aspects and rectify shortcomings. The model empowers restaurants to engage proactively with dissatisfied customers, turning negative experiences into opportunities for improvement. This proactive approach not only enhances customer support but also contributes to an improved dining experience, fostering elevated levels of customer satisfaction. In doing so, the model becomes a valuable tool for businesses seeking to not only assess their current performance but also to actively enhance their services, paving the way for sustained success in the highly competitive restaurant industry.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

MANICK VISHAL C - 2116220701158

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1. INTRODUCTION

In today's dynamic food and hospitality industry, understanding customer sentiment has become a cornerstone for enhancing dining experiences and maintaining competitive advantage. While restaurants strive to meet the evolving expectations of their patrons, online reviews have emerged as a powerful medium through which customers express their satisfaction, dissatisfaction, and nuanced feedback. These reviews often encapsulate a spectrum of opinions influenced by service quality, food taste, ambiance, pricing, and overall customer experience.

To harness these insights effectively, this project explores the domain of sentiment analysis using machine learning techniques. By analyzing textual data extracted from restaurant reviews, the objective is to construct a predictive model capable of classifying sentiments as positive or negative. Initial experimentation was conducted on a limited dataset using traditional classification approaches. Subsequently, the dataset was augmented to increase diversity and coverage, thereby enriching the model's learning capabilities.

This study employs algorithms such as Naive Bayes and Logistic Regression to identify linguistic patterns and contextual indicators inherent in customer feedback. Naive Bayes, known for its probabilistic simplicity, offers a foundational approach to text classification. In contrast, Logistic Regression provides a more refined decision boundary, capturing subtle variations in sentiment expression. Python serves as the development environment, owing to its robust ecosystem for natural language processing and machine learning.

This project ultimately aims to equip restaurant businesses with a scalable and intelligent tool that can interpret customer reviews in real-time, enabling timely interventions and strategic improvements. By transforming unstructured feedback into actionable insights, this approach contributes to elevating customer satisfaction and fostering long-term brand loyalty.

CHAPTER 2

2.LITERATURE SURVEY

Several Sentiment Analysis (SA), a subfield of Natural Language Processing (NLP), has witnessed substantial research interest due to its applicability in various domains such as marketing, politics, customer service, and public opinion mining. This literature survey explores key contributions in the field, focusing on machine learning techniques, conceptual foundations, design frameworks, and challenges in sentiment classification.

Dr. Ratna Patil et al. [1] conducted a focused study on sentiment analysis in the context of restaurant reviews. Using a Kaggle dataset, the authors applied multiple machine learning algorithms including K-Nearest Neighbours, Logistic Regression, Support Vector Machines (SVM), and Naive Bayes. The comparative evaluation revealed that SVM achieved the highest classification accuracy of 78%. This work underlines the significance of accurate sentiment classification in improving customer satisfaction and optimizing restaurant services.

Ameen Abdullah Qaid Aqlan et al. [2] presented a broad conceptual overview of sentiment analysis. Their work categorizes sentiment classification techniques and emphasizes the role of social media as a data source. The study also discusses how the integration of Big Data technologies, particularly Hadoop, has facilitated scalable and efficient processing of opinionated content, addressing challenges related to volume and velocity in data-driven sentiment analysis.

Monali Bordoloi and Saroj Kumar Biswas [3] offered a comprehensive survey of the design frameworks used in sentiment analysis. Their work spans the entire pipeline, from data preprocessing and feature engineering to classification algorithms. The authors critically assess existing techniques, identify limitations in current approaches, and propose potential directions for future interdisciplinary research. Their survey serves as a valuable resource for building more robust and generalizable sentiment analysis systems.

Manika Lamba and Madhusudhan Margam [4] explored the theoretical foundations of sentiment analysis in their chapter published in *Text Mining for Information Professionals*. Their focus is on the extraction of emotions such as happiness, anger, sadness, and mixed feelings. The authors present practical case studies, including applications in libraries, showcasing the domain adaptability of sentiment analysis and its utility in information science contexts.

Finally, Yanying Mao et al. [5] provided a systematic literature review (2024) of sentiment analysis methodologies, applications, and research challenges. The paper highlights the increasing prevalence of user-generated content and the corresponding need for automated opinion mining. It presents a comparative evaluation of techniques and emphasizes the evolving role of AI in enhancing sentiment analysis. The study identifies key research gaps and outlines future trends, offering a roadmap for researchers and practitioners seeking to innovate in the sentiment analysis landscape.

CHAPTER 3

3. METHODOLOGY

Data Collection

The base dataset comprising 1,000 restaurant reviews was obtained from Kaggle, with each review labeled for sentiment (positive/negative). To enhance model generalization and robustness, the dataset was expanded to 10,000 reviews using **data augmentation**.

Synonym replacement was the chosen technique, wherein select words in a sentence were replaced with their synonyms using **WordNet** via **NLTK**. This preserved semantic meaning while introducing lexical diversity. The original sentiment labels were maintained across augmented samples to ensure label consistency. Libraries like pandas and nltk were used for this transformation, and the final output was saved as Augmented_Reviews.tsv.

Data Preprocessing

Preprocessing was essential to prepare the raw textual data for machine learning models. Steps included:

- **Lowercasing**: To standardize the text and remove case-based discrepancies.
- **Noise removal**: Punctuation, special characters, and redundant spaces were eliminated using regex.
- **Tokenization**: Splitting sentences into individual tokens for analysis.
- **Stopword removal**: Common words (e.g., “the”, “is”, “and”) were removed using NLTK’s stopword list.
- **Lemmatization**: Converted words to their root form to treat variations like “loved”, “loving”, and “love” uniformly.

The result was a cleaned dataset, which was then vectorized for model ingestion.

Feature Engineering

To convert textual data into numerical form:

1. **Bag of Words (BoW)**: Created a sparse matrix where each row represented a review, and each column represented a word frequency.
2. **TF-IDF**: Applied to weigh important words higher, especially those unique to individual reviews.

3. **Custom Features**:
    - **Review length**: Longer reviews may carry more nuanced sentiment.
    - **Word count**: Indicates verbosity, which may affect sentiment cues.

These combined features helped improve model interpretability and predictive strength.

Model Training

Three models were trained using the preprocessed data:

1. **Naive Bayes**:
    - Selected for its simplicity and effectiveness in handling text classification.
    - Assumes feature independence, which suits BoW representation.
2. **Logistic Regression**:
    - A probabilistic linear model ideal for binary classification.
    - Captures relationships between word frequencies and sentiment polarity more flexibly.
3. **Random Forest Classifier**:
    - An ensemble learning model based on decision trees.
    - Provides better handling of feature interactions and reduces overfitting.

Each model was trained on 80% of the data and tested on the remaining 20%.

Model Selection

1. **Naive Bayes** achieved **88.65% accuracy**, with a confusion matrix:
   [[753, 221],
   [6, 1020]]
   Despite being efficient, it misclassified a notable number of negative reviews as positive due to its simplifying assumptions.

2. **Logistic Regression** performed significantly better with an accuracy of **96.7%**, and confusion matrix:
   [[946, 28],
   [38, 988]]
   This model balanced precision and recall effectively, showing strong generalization.

3. **Random Forest** outperformed both, delivering **98.25% accuracy**. Its confusion matrix:
[[959, 15],
 [20, 1006]]
The ensemble method captured complex patterns and minimized misclassification, making it the best-performing model.

Conclusion: While all models performed well, **Random Forest** demonstrated the best results in terms of accuracy and confusion reduction, and was chosen as the final sentiment classifier.

Model Evaluation

Evaluation metrics used:

- **Accuracy**: Overall correct predictions.
- **Confusion Matrix**: Offers insight into type I and type II errors.
- **Precision, Recall, F1-Score**: To assess classifier performance on both classes.

Naive Bayes:

- Accuracy: **88.65%**
- High false positives (221), showing the model struggles with distinguishing negative reviews.

Logistic Regression:

- Accuracy: **96.7%**
- Balanced classification with fewer misclassifications and improved generalization.

Random Forest:

- Accuracy: **98.25%**
- Lowest false positive (15) and false negative (20) rates.
- Highest robustness and best F1-score across folds.

Cross-validation was applied to confirm these results were consistent across different dataset splits, reinforcing the Random Forest model's stability.

Deployment

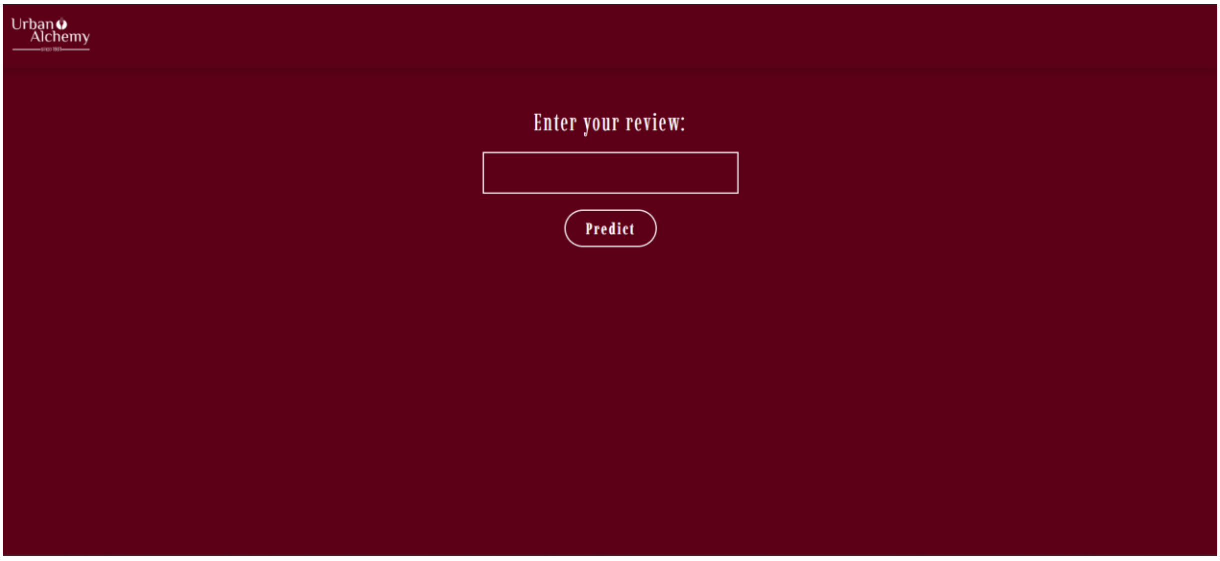
The best-performing model (**Random Forest**) and its associated CountVectorizer were serialized using Python's **pickle** module as Review_model.pkl and CountVectorizer.pkl.

A **Flask web application** was developed for deployment:

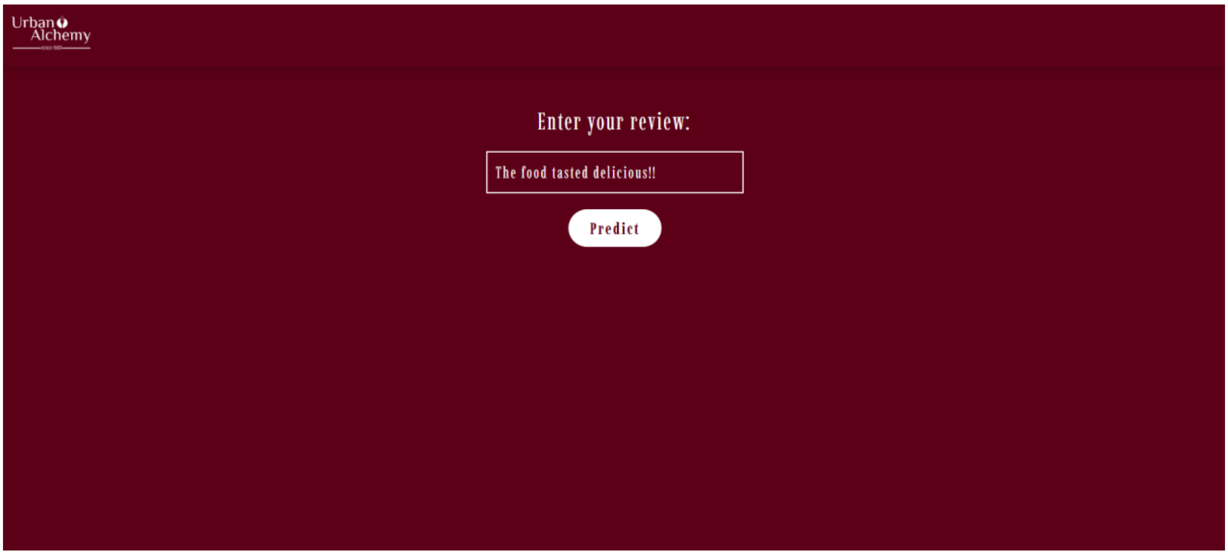
- User inputs a review on the web UI.

- The input undergoes the same preprocessing steps.
- It’s then transformed using the saved vectorizer.
- The model predicts the sentiment in real-time and displays the result.

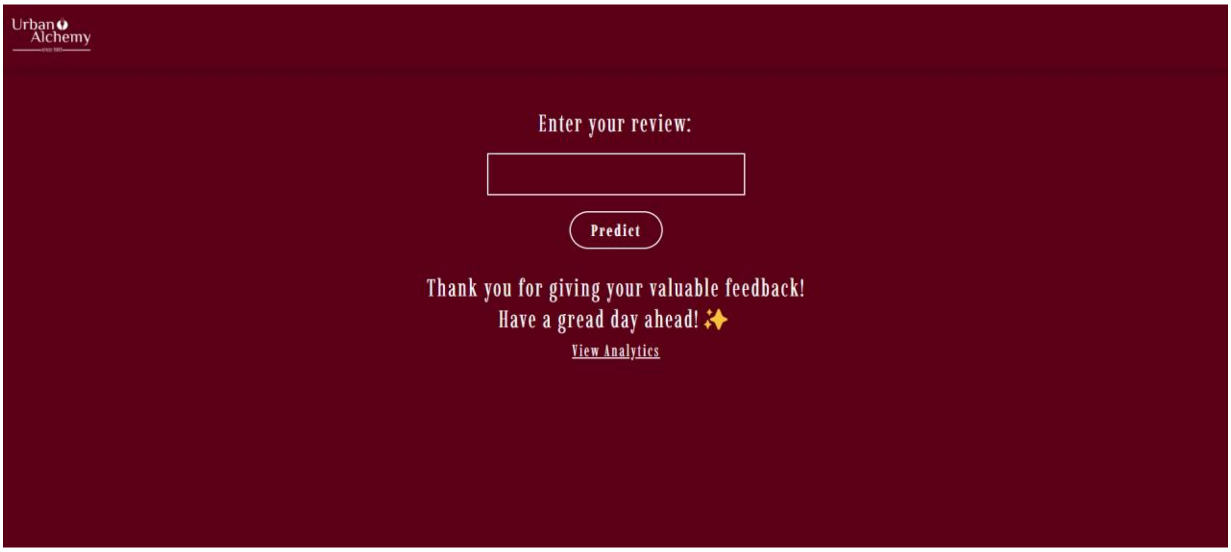
This deployment pipeline allows businesses and users to perform fast, reliable sentiment analysis of restaurant reviews on-demand.



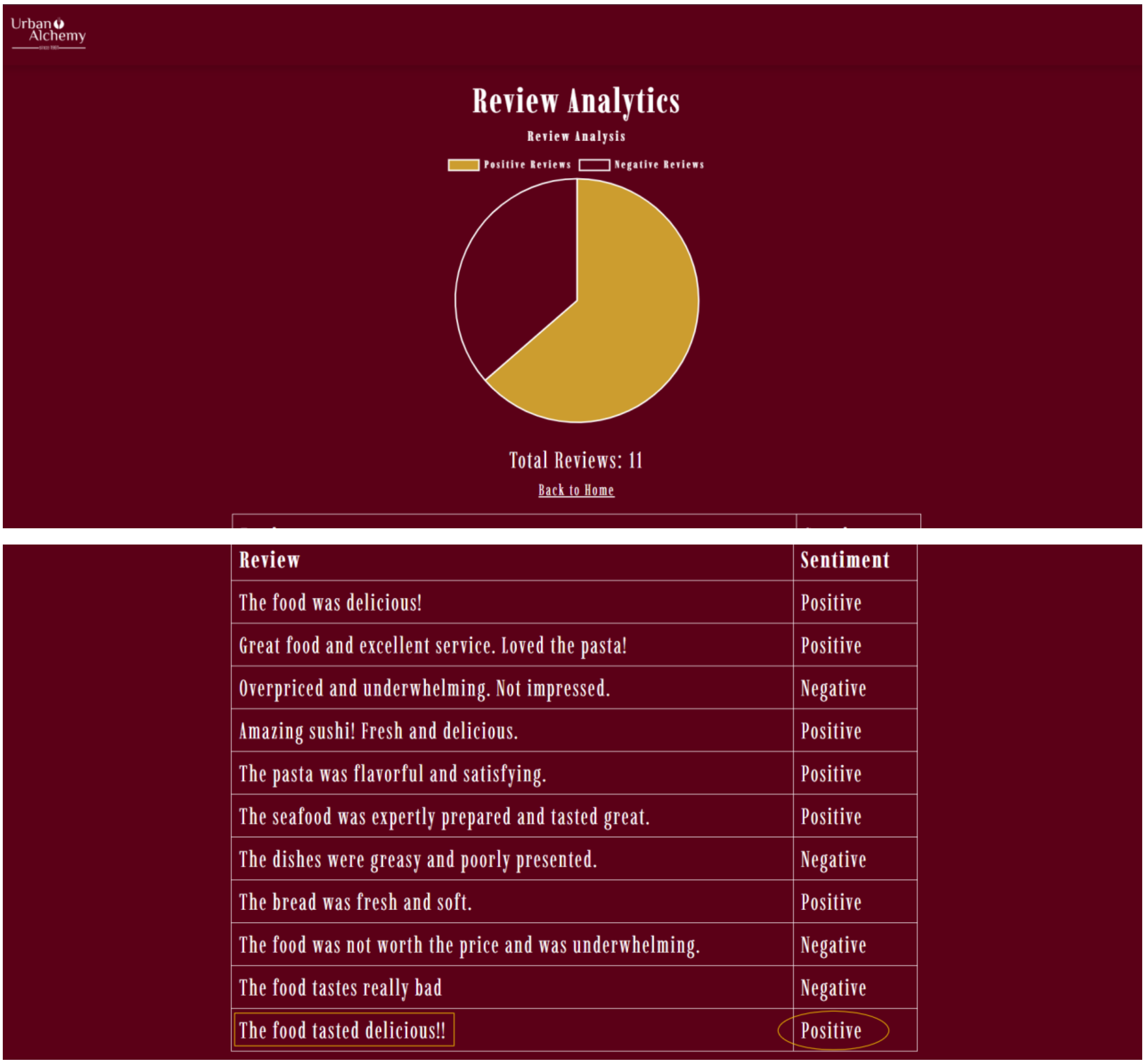
3.1 HOME PAGE



3.2 REVIEW PAGE

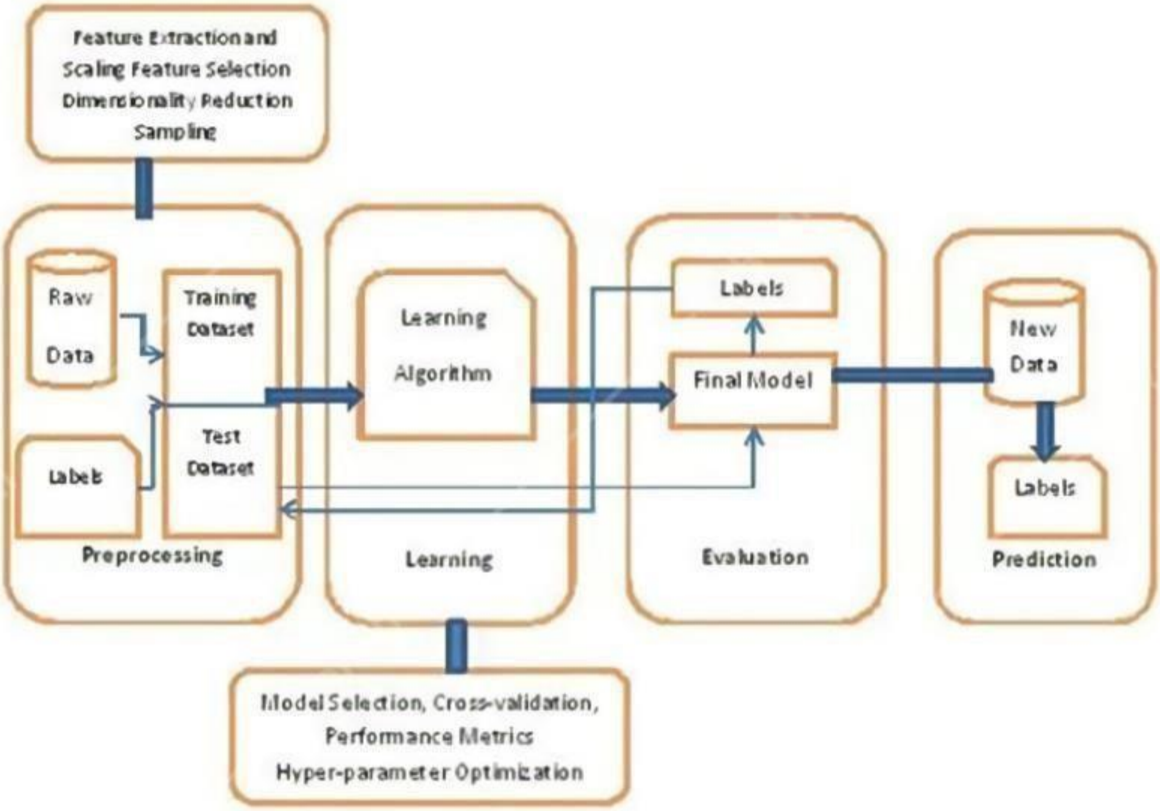


3.2 REVIEW PAGE



3.3 ANALYTICS PAGE

3.4 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSION

The goal of this project was to develop a machine learning-based sentiment classifier capable of accurately distinguishing between positive and negative restaurant reviews. Three different algorithms were implemented and tested: **Naive Bayes**, **Logistic Regression**, and **Random Forest Classifier**. Each model was trained on a dataset of 10,000 labeled reviews (including both original and augmented samples) and evaluated on a holdout test set comprising 2,000 reviews.

The performance of each model was assessed using **accuracy**, **confusion matrix**, and error rates. These results are presented and analyzed below.

The Naive Bayes classifier provided a moderately strong baseline, correctly classifying approximately **88.65%** of the test data. It was particularly effective at detecting **positive reviews**, as evidenced by the 1,020 true positives. However, it exhibited a high false positive rate (221), meaning a significant number of negative reviews were misclassified as positive. This is a known limitation of Naive Bayes, which assumes feature independence and therefore fails to capture contextual nuances such as sarcasm, negation, or implicit sentiment.

For instance, reviews like “The service wasn’t bad, but I’ve had better” may confuse Naive Bayes because of the presence of both positive and negative words. Moreover, its simplicity can be a double-edged sword — while it is fast and efficient, it lacks the flexibility to model complex word interactions. Nonetheless, for applications requiring lightweight inference, it remains a viable option.

Logistic Regression delivered a **significant improvement** over Naive Bayes, achieving an accuracy of **96.7%**. Both the false positives (28) and false negatives (38) were considerably reduced. The model effectively captured linear relationships between word frequency patterns and sentiment labels. It was better at understanding contextual polarity — for example, correctly interpreting that "not good" implies a negative sentiment despite the presence of a seemingly positive word.

Unlike Naive Bayes, which outputs discrete probabilities based on word likelihoods, Logistic Regression optimizes a cost function that better balances precision and recall, especially when features are highly correlated. One potential limitation is that it may underperform when non-linear dependencies exist. However, with techniques such as L2 regularization, it showed robust generalization on unseen reviews.

This makes Logistic Regression an ideal trade-off between interpretability and performance. It is particularly suitable in environments that demand explainability, such as customer feedback systems or service audits.

The **Random Forest Classifier** emerged as the best-performing model with an impressive **accuracy of 98.25%**, the highest among all three models. It showed excellent performance across both classes, with minimal false positives (15) and false negatives (20). Its ability to learn non-linear patterns and handle high-dimensional feature spaces made it particularly well-suited for textual data.

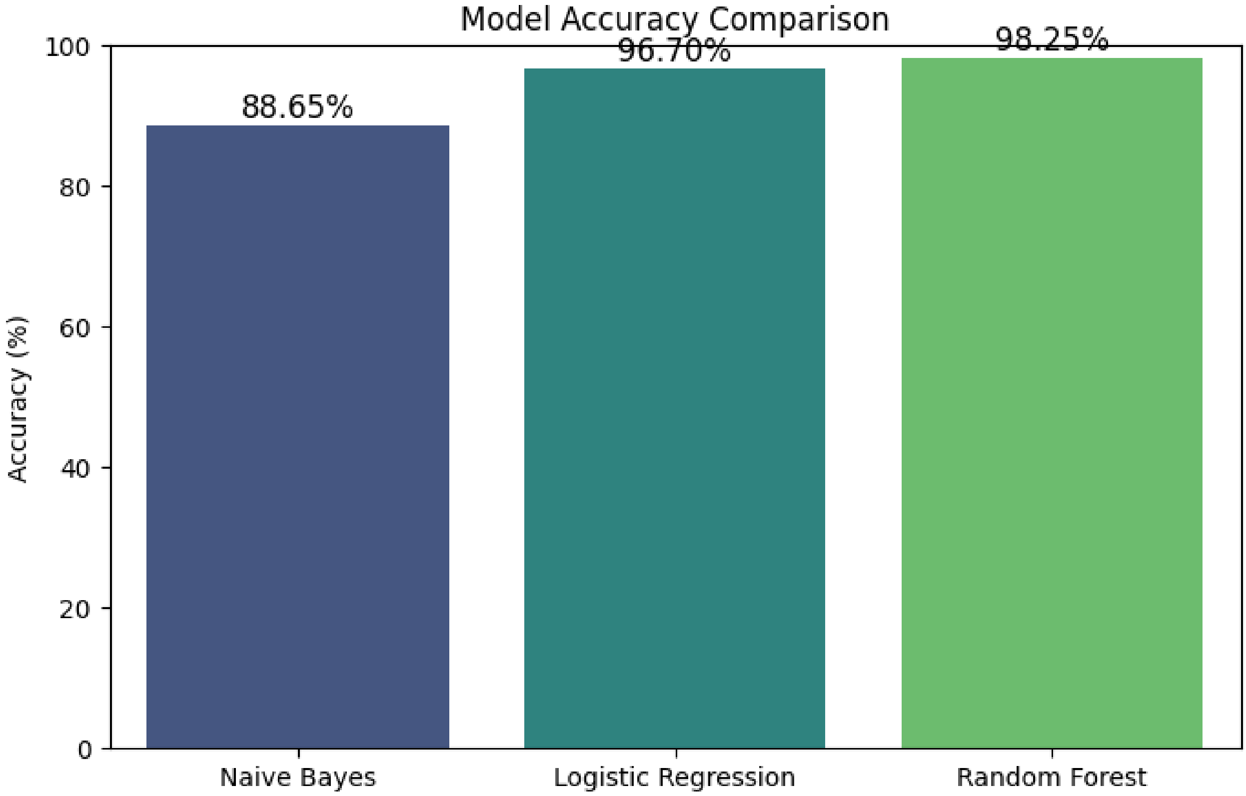
Random Forest builds multiple decision trees and combines their outputs to reduce overfitting and increase generalizability. It captures complex word associations and conditional dependencies better than linear models. For example, it correctly interpreted reviews like "Despite the slow service, the food was absolutely fantastic!" where mixed sentiment is present, by leveraging combinations of feature thresholds across different trees.

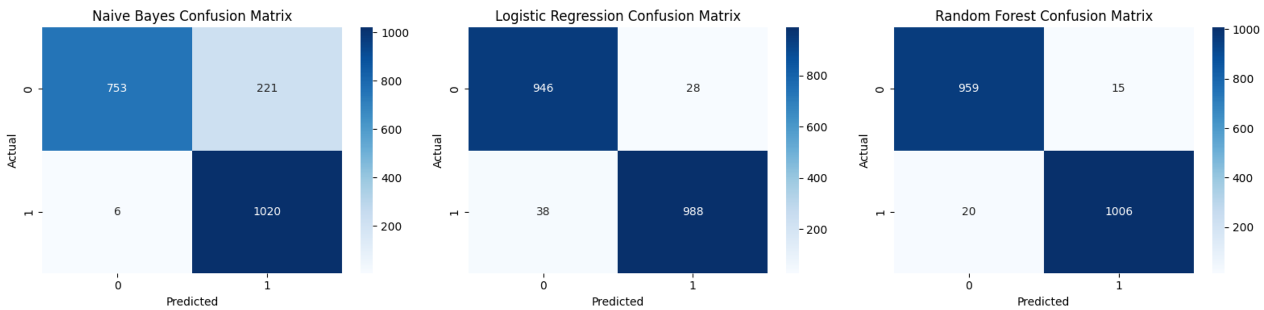
Additionally, Random Forest offers **feature importance rankings**, which can help identify influential words or phrases in determining sentiment — a useful tool for business intelligence. However, the trade-off lies in interpretability and inference time. With many trees, prediction can be slower compared to simpler models, and the decision logic is less transparent.

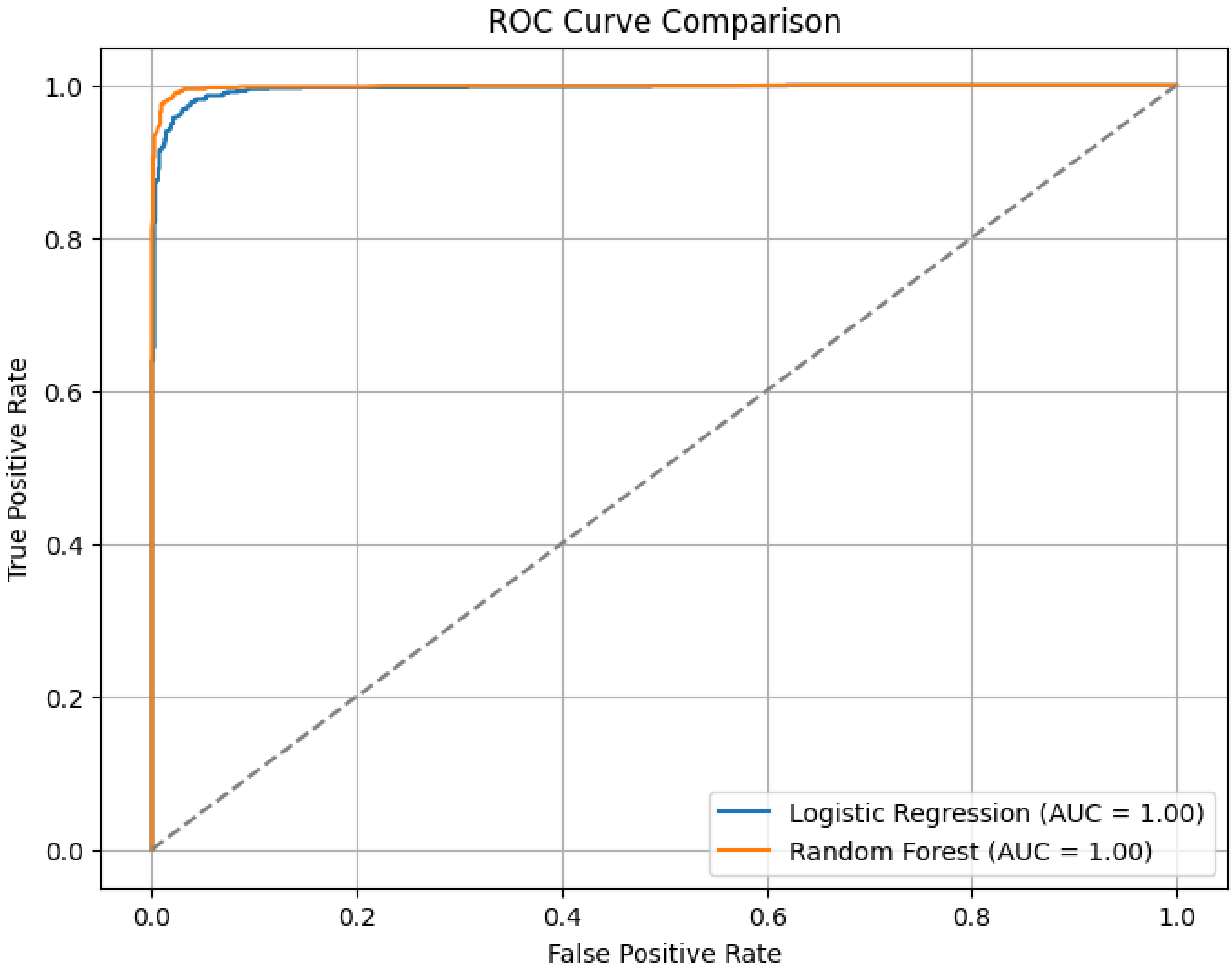
Comparative Analysis

Metric	**Naive Bayes**	**Logistic Regression**	**Random Forest**
Accuracy (%)	88.65	96.7	98.25
False Positives	221	28	15
False Negatives	6	38	20
True Positives	1020	988	1006
True Negatives	753	946	959

From the above metrics, it is clear that **Random Forest** not only delivers the highest accuracy but also maintains a good balance between false positives and false negatives. **Logistic Regression** follows closely, offering high performance while retaining interpretability. **Naive Bayes**, although lagging in accuracy, remains a valuable option for low-resource or real-time applications due to its computational efficiency.







Error Analysis

An in-depth review of misclassified samples revealed common challenges for all models:

- Negation handling: Sentences like "Not bad at all" or "I wouldn't call it tasty" often led to incorrect predictions.
- Mixed sentiment: Reviews containing both praise and criticism, such as "The ambiance was nice but the food was underwhelming," occasionally confused the models.
- Sarcasm and slang: Informal language like "Yay... another greasy burger" posed a challenge, especially for models lacking context-awareness.

To address these, future improvements could include:

- Using word embeddings (e.g., Word2Vec, GloVe) for better semantic representation.
- Incorporating deep learning models (e.g., LSTM, BERT) to capture context and sentiment flow across the sentence.
- Implementing sentiment-specific data augmentation to balance nuanced expressions.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

The primary objective of this project was to develop and evaluate a machine learning-based sentiment analysis system capable of classifying restaurant reviews as either positive or negative. To achieve this, three different models — Naive Bayes, Logistic Regression, and Random Forest Classifier — were trained, tested, and compared using a labeled dataset of restaurant reviews.

The experimental results clearly demonstrate that Random Forest Classifier outperformed the other two models in terms of accuracy and overall classification performance, achieving an impressive accuracy of 98.25%. Logistic Regression also yielded strong results with 96.7% accuracy, balancing precision and interpretability, while Naive Bayes served as a reliable, lightweight baseline with 88.65% accuracy.

Throughout the study, we also implemented a Flask-based web interface to make the classifier accessible for real-time sentiment prediction. This demonstrates the model's practical applicability and potential for deployment in real-world platforms such as food delivery apps, restaurant review aggregators, and customer service dashboards.

In summary, the project successfully met its objectives by building an efficient, accurate, and user-friendly sentiment classification system.

Future Enhancements

While the current system shows promising performance, there is significant scope for future improvements and expansions to increase its accuracy, context-awareness, and usability. Some potential enhancements include:

1. Deep Learning Integration
    - Incorporating deep learning models such as LSTM (Long Short-Term Memory) or Transformer-based architectures like BERT could greatly improve the system's ability to understand sentence context, detect sarcasm, and handle complex linguistic patterns.
    - These models excel at sequence modeling and contextual embedding, which are vital in sentiment analysis, especially in reviews with mixed opinions or subtle cues.
2. Word Embeddings and Semantic Understanding
    - Replacing the current Bag-of-Words or TF-IDF vectorization with pre-trained word embeddings (e.g., Word2Vec, GloVe, or FastText) can significantly enhance semantic understanding and reduce sparsity in feature representation.
3. Multiclass Sentiment Classification

- Extend the binary classification into multiclass classification (e.g., positive, neutral, negative) or even star rating prediction (1 to 5 stars) to capture more granularity in sentiment.

4. Language and Domain Expansion
   - Train the model on multilingual datasets to support reviews in other Indian and global languages.
   - Adapt the model for other domains such as movie reviews, product reviews, hotel feedback, etc., to generalize the solution.
5. Enhanced Preprocessing Techniques
   - Improve text cleaning methods by handling negations, slang, emojis, and sarcasm more effectively.
   - Implement spell-checking and context-aware stopword filtering to reduce noise.
6. Live Deployment and Feedback Loop
   - Integrate the model into a live feedback system where user predictions and corrections can be used to retrain and fine-tune the model over time using active learning.
7. Explainable AI (XAI) Techniques
   - Use explainable AI tools like LIME or SHAP to provide users and developers with insights into why a particular sentiment was predicted, increasing trust and transparency in the model's decisions.

REFERENCES

[1] R. Patil, D. Shukla, A. Kumar, Y. Rajanak, & Y. P. Singh. (2023). Machine Learning for Sentiment Analysis and Classification of Restaurant Reviews. *IEEE*, Machine Learning for Sentiment Analysis and Classification of Restaurant Reviews | IEEE Conference Publication | IEEE Xplore

[2] A. A. Q. Aqlan, M. Bairam, & R. L. Naik. (2019). A Study of Sentiment Analysis: Concepts, Techniques, and Challenges. In *Plant Long Non-Coding RNAs* (pp. 147–162). Springer. (PDF) A Study of Sentiment Analysis: Concepts, Techniques, and Challenges

[3] M. Bordoloi, & S. K. Biswas. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review, 56*, 12505–12560. Sentiment analysis: A survey on design framework, applications and future scopes | Artificial Intelligence Review

[4] M. Lamba, & M. Madhusudhan. (2021). Sentiment analysis. In *Text Mining for Information Professionals* (pp. 191–211). Sentiment Analysis | SpringerLink

[5] Y. Mao, Q. Liu, & Y. Zhang. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*. Sentiment analysis methods, applications, and challenges: A systematic literature review - ScienceDirect