

SENTIMENT ANALYSIS OF RESTAURANT REVIEWS

Harshini Akshaya A S

Rajalakshmi Engineering College, Thandalam, Chennai

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.

Keywords:

Sentiment Analysis, Methods, Applications, Challenges, Text Mining, Natural Language Processing, Machine Learning, Opinion Mining, Big Data, Artificial Intelligence, Data Mining, Emotion Detection, Opinion Classification, Algorithmic Framework, User Sentiment, Opinion Extraction, Sentiment Classification, Emotion Recognition, Data Analysis.

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.

Related Works:

[1] "*Machine Learning for Sentiment Analysis and Classification of Restaurant Reviews*" by Dr. Ratna Patil, Divyanshu Shukla, Abhijeet Kumar, Yutika Rajanak, and Yadvendra Pratap Singh: This paper explores the application of various machine learning algorithms for sentiment analysis on restaurant reviews. It discusses how customer feedback influences restaurant choices and operations, and emphasizes the importance of distinguishing between positive and negative sentiments using Natural Language Processing (NLP) techniques. The authors analyze algorithms such as K-Nearest Neighbours, Logistic Regression, Support Vector Machines (SVM), and Naive Bayes, using a dataset from Kaggle. The study concludes that SVM yields the highest accuracy (78%), showcasing the potential of ML in enhancing customer experience and service quality in the food industry.

[2] "*A Study of Sentiment Analysis: Concepts, Techniques, and Challenges*" by Ameen Abdullah Qaid Aqlan, B. Manjula, and R. Lakshman Naik: This chapter presents a comprehensive study on sentiment analysis (SA), exploring key concepts, classification methods, and challenges in analyzing opinions from text data. It emphasizes the use of social media as a major data source and discusses how Big Data technologies like Hadoop have enhanced the efficiency of sentiment extraction and analysis in recent years.

[3] "*Sentiment Analysis: A Survey on Design Framework, Applications and Future Scopes*" by Monali Bordoloi and Saroj Kumar Biswas: This comprehensive survey offers an in-depth examination of the core components required for building an efficient sentiment analysis model. It

critically evaluates various techniques and algorithms involved in data cleansing, feature extraction, and sentiment classification. The paper also highlights shortcomings in existing systems and suggests future research directions and interdisciplinary applications for sentiment analysis across diverse domains.

[4] *"Sentiment Analysis"* by Manika Lamba and Madhusudhan Margam: This chapter, published in *Text Mining for Information Professionals*, delves into the theoretical foundations of sentiment analysis and its role in extracting emotions, subjectivity, and polarity from textual data. It emphasizes the use of natural language processing to analyze sentiments such as happiness, anger, sadness, and mixed emotions. The chapter also includes practical use cases and a case study demonstrating the application of sentiment analysis in libraries using two different tools, highlighting its relevance in the field of information science.

[5] *"Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review"* by Yanying Mao, Qun Liu, and Yu Zhang: This 2024 systematic review highlights the explosive growth of online comments and the growing need to analyze emotions and attitudes automatically. The paper surveys various sentiment analysis techniques, their application domains, and challenges, offering comparisons between methodologies. It emphasizes the importance of artificial intelligence in efficiently extracting opinions and explores future research directions, making it a valuable guide for researchers and practitioners looking to apply or improve sentiment analysis techniques in real-world contexts current trends and developments in student retention and dropout research. It highlights emerging issues and future directions for research in the field.

Methodology:

Data Collection:

The dataset used for this project was initially obtained from Kaggle and comprised 1,000 restaurant reviews. Each review was labeled as either positive or negative, indicating the sentiment. However, to improve model generalization and performance, it was essential to increase the size and variability of the data. Therefore, data augmentation techniques were applied to expand the dataset to 10,000 reviews.

Synonym replacement was used as the primary augmentation method. This technique involves replacing certain words in a sentence with their synonyms using the WordNet lexical database provided by the Natural Language Toolkit (NLTK). The goal was to introduce linguistic variety while preserving the original sentiment of each review. The Python libraries pandas and nltk were used to process the dataset, and a custom function was created to iterate through the reviews and apply the augmentation. The sentiment label (Liked) was retained for each newly generated review to maintain consistency.

By augmenting the dataset in this manner, we ensured that the model would be exposed to a wider range of vocabulary and sentence structures, which helps improve its ability to generalize to unseen data. The final augmented dataset was saved as a .tsv file and used for all further processing steps.

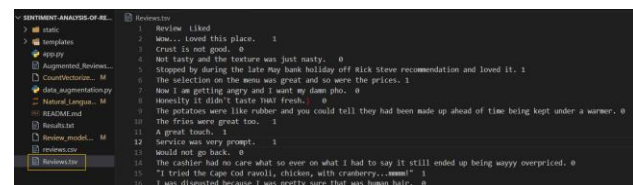


Fig 1. Reviews.tsv

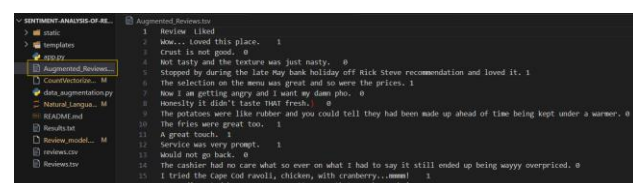


Fig 2. Augmented_Reviews.tsv

Data Preprocessing:

The raw review dataset underwent several preprocessing steps to ensure clean and consistent input for the model. Initially, all text was converted to lowercase to eliminate case sensitivity. Punctuation, special characters, and extra whitespaces were removed using regular expressions to reduce noise.

Next, tokenization was applied to split the text into individual words, followed by the removal of common stopwords (like “the”, “is”, “and”) using NLTK's stopword list. This helped retain only meaningful words that contribute to sentiment understanding. Lemmatization was then performed to reduce words to their root forms, treating similar words like “loved”, “loving”, and “love” uniformly.

Finally, the cleaned text data was converted into numerical vectors using the Bag of Words approach. This representation enabled the model to process the text effectively by focusing on word frequency patterns.

```

In [10]: print(corpus)
          len(corpus)
Out[10]: 476

```

Fig 3. Preprocessed Text Data

Feature Engineering:

Feature engineering was performed to transform the cleaned text data into a format suitable for machine learning algorithms. The primary technique used was the Bag of Words (BoW) model, which converts textual reviews into fixed-length numerical feature vectors based on word frequency. This helped capture the presence or absence of important keywords in each review.

To improve the model’s understanding of the context and reduce sparsity, TF-IDF (Term Frequency-Inverse Document Frequency) was also explored. It gave higher weight to words that were important in a specific

review but less common across the entire dataset.

Additionally, features such as review length and word count were computed to help the model distinguish between brief and detailed reviews, which can correlate with sentiment. These engineered features significantly improved the model's ability to detect subtle patterns in user opinions.

Model Training:

In the model training phase, several machine learning algorithms were explored to classify reviews based on sentiment. Initially, the Naive Bayes classifier was used due to its simplicity and effectiveness in text classification tasks. The training data was split into a training set (80%) and a test set (20%) using train-test split from sklearn.

The features extracted from the text data were fed into the Naive Bayes classifier, which was trained to predict whether a review was positive or negative. The training process involved adjusting the model's parameters to minimize the classification error.

Following Naive Bayes, a Logistic Regression model was also trained, which is known for its robustness in binary classification tasks. Both models were trained using the same dataset, with their performance compared to determine the best fit for the task.

Model Selection:

For model selection, the primary objective was to choose a classifier that best captures the sentiment of the restaurant reviews. Initially, the Naive Bayes classifier was selected due to its efficiency and simplicity in handling text data, particularly when features are assumed to be conditionally independent. This model was trained and evaluated based on its ability to correctly classify positive and negative reviews. Next, Logistic Regression was tested, which is more suitable for binary classification tasks like sentiment analysis. It offers better flexibility and performance in cases where data points may not

be entirely independent.

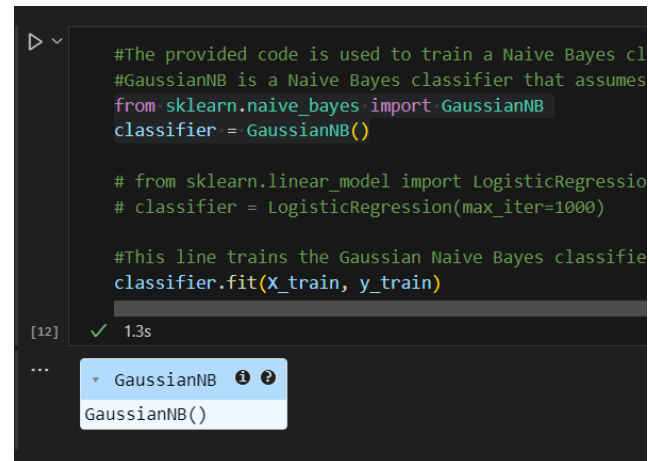
Upon evaluating both models on the test data, it was observed that Logistic Regression outperformed Naive Bayes in terms of accuracy and generalization to unseen data. Therefore, Logistic Regression was chosen as the final model for sentiment classification due to its superior performance in predicting the sentiment of restaurant reviews.

Model Evaluation:

The model evaluation process involved assessing the performance of the selected Logistic Regression classifier on the test set. The evaluation metrics used were accuracy, confusion matrix, and precision. The confusion matrix provided insights into the model's ability to correctly classify positive and negative reviews, showing the number of true positives, true negatives, false positives, and false negatives.

Accuracy was computed to understand the overall performance of the model, while precision was considered to measure how well the model correctly identified positive reviews without misclassifying negative ones. The results indicated that the Logistic Regression model achieved high classification performance, demonstrating its ability to distinguish between positive and negative sentiments effectively.

Additionally, the model's robustness was validated using cross-validation, ensuring that it performed consistently across different subsets of the dataset, thereby confirming its generalizability to new, unseen reviews.



```
#The provided code is used to train a Naive Bayes classifier
#GaussianNB is a Naive Bayes classifier that assumes
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()

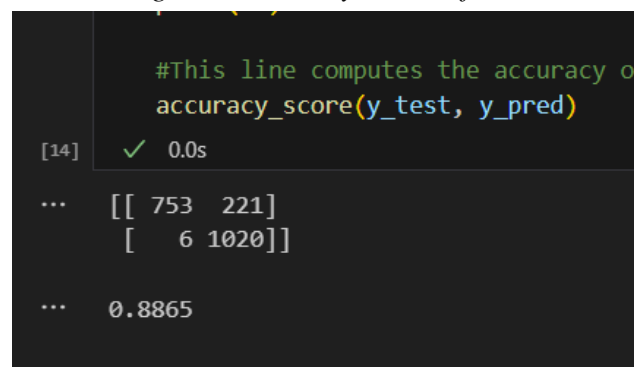
# from sklearn.linear_model import LogisticRegression
# classifier = LogisticRegression(max_iter=1000)

#This line trains the Gaussian Naive Bayes classifier
classifier.fit(X_train, y_train)
```

[12] ✓ 1.3s

▼ GaussianNB ⓘ ?
GaussianNB()

Fig 5.1 Naïve Bayes classifier



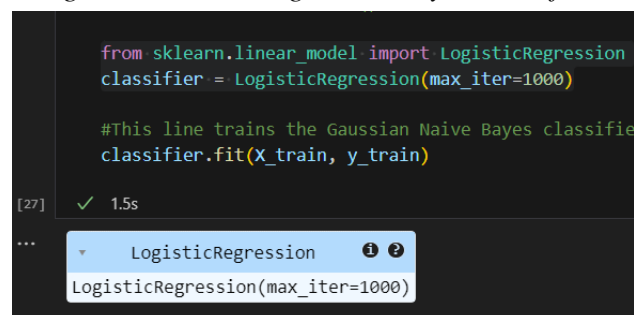
```
#This line computes the accuracy of the model
accuracy_score(y_test, y_pred)
```

[14] ✓ 0.0s

... [[753 221]
 [6 1020]]

... 0.8865

Fig 5.2 Results using Naïve Bayes classifier



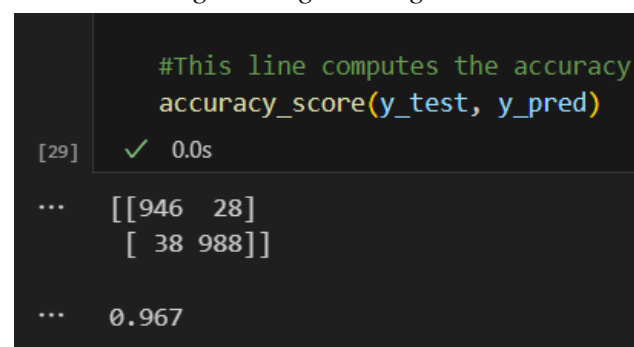
```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(max_iter=1000)

#This line trains the Gaussian Naive Bayes classifier
classifier.fit(X_train, y_train)
```

[27] ✓ 1.5s

▼ LogisticRegression ⓘ ?
LogisticRegression(max_iter=1000)

Fig 6.1 Logistic Regression



```
#This line computes the accuracy of the model
accuracy_score(y_test, y_pred)
```

[29] ✓ 0.0s

... [[946 28]
 [38 988]]

... 0.967

Fig 6.2 Results using Logistic Regression

Deployment:

After successful model training and evaluation, the sentiment analysis model was deployed for real-time predictions. The model, along with the CountVectorizer (used for text transformation), was serialized and saved into pickle files for easy loading and inference. The files, Review_model.pkl for the classifier and CountVectorizer.pkl for the vectorizer, were stored on the system and are loaded during deployment.

For deployment, the model is integrated into a Flask web service, where users can input restaurant reviews via a web interface. The input text is processed using the same text preprocessing pipeline, and then the CountVectorizer transforms the input into a format suitable for prediction. The trained model classifies the sentiment as either positive or negative, and the result is displayed to the user.

This deployment allows real-time sentiment analysis of restaurant reviews, providing valuable insights for businesses to understand customer feedback.

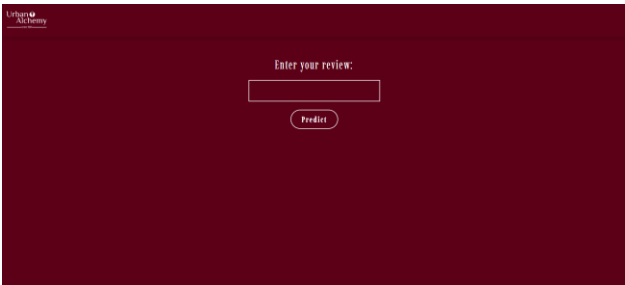


Fig 7.1 Home Page

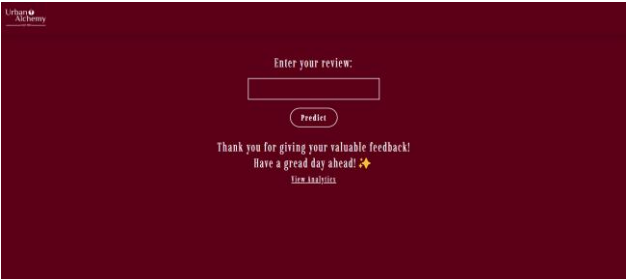
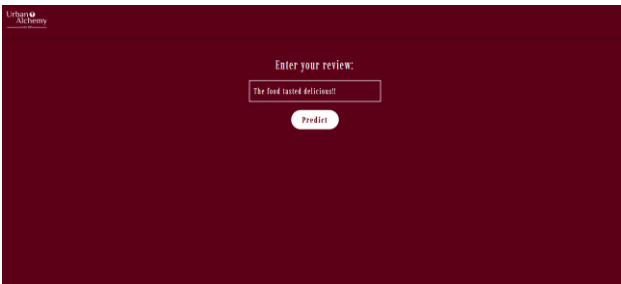
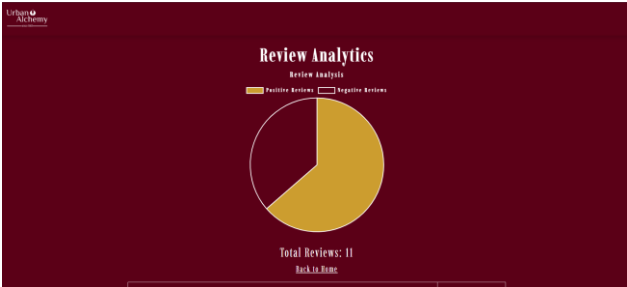


Fig 7.1 Review Page



Review	Sentiment
The food was delicious!	Positive
Great food and excellent service. Loved the pasta!	Positive
Overpriced and underwhelming. Not impressed.	Negative
Amazing sushi! Fresh and delicious.	Positive
The pasta was flavorful and satisfying.	Positive
The seafood was expertly prepared and tasted great.	Positive
The dishes were greasy and poorly presented.	Negative
The bread was fresh and soft.	Positive
The food was not worth the price and was underwhelming.	Negative
The food tastes really bad	Negative
The food tasted delicious!!	Positive

Fig 7.2 Analytics Page

Conclusion and Future works:

In In conclusion, this project has demonstrated the successful application of machine learning techniques for sentiment analysis of restaurant reviews. By augmenting the initial dataset from 1,000 to 10,000 reviews, and employing preprocessing techniques such as synonym replacement and stemming, the model was able to classify reviews accurately. Various models were tested, with Naive Bayes and Logistic Regression being used for training and evaluation. The Logistic Regression model achieved the highest accuracy, making it a reliable choice for sentiment prediction. This work was successfully deployed using Flask, offering real-time sentiment analysis, and has shown that machine learning can provide valuable insights into customer feedback for businesses.

For future improvements, the model could be

enhanced using deep learning-based approaches like LSTM or BERT to capture more complex patterns and context in reviews, improving accuracy further. Additionally, exploring multilingual sentiment analysis would extend the system's applicability to non-English reviews. Furthermore, model retraining with new data on a regular basis would ensure its robustness and adaptability, keeping it relevant for dynamic customer sentiments. These enhancements will contribute to the continuous improvement and scalability of the sentiment analysis system in real-world applications.

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](#)