Zomato Restaurant Review Sentiment Analysis

An End-to-End NLP & Machine Learning Project

Prepared by: Raj Bhagwan Sonar

1. Introduction

Customer reviews play a critical role in shaping the reputation of restaurants on online food delivery platforms. Platforms such as Zomato host thousands of reviews every day, making manual analysis impractical. This project focuses on automatically analyzing customer reviews using Natural Language Processing (NLP) and Machine Learning techniques to classify customer sentiment and extract meaningful business insights.

2. Business Problem & Objective

Restaurants often rely only on numeric ratings to evaluate performance. However, ratings alone do not explain why customers are satisfied or dissatisfied. The business problem addressed in this project is to bridge this gap by analyzing textual feedback and converting it into structured insights.

Project Objectives:

1. Automatically classify customer reviews as positive or negative.
2. Identify key factors influencing customer sentiment such as food and service.
3. Enable restaurant-level performance comparison.
4. Build a deployable sentiment prediction system.

3. Dataset Description

The project uses two datasets: a restaurant metadata dataset and a customer reviews dataset. The reviews dataset contains approximately 10,000 records, each consisting of a restaurant name, a textual review, and a customer-provided rating. This dataset is well-suited for both exploratory analysis and supervised machine learning.

4. Data Cleaning & Preparation

Before performing analysis and modeling, the dataset was carefully cleaned to ensure data quality. Ratings were converted to numeric format, and invalid or missing ratings were removed. Neutral reviews were excluded to maintain a clear binary sentiment classification problem. Textual missing values were handled safely to avoid runtime errors during preprocessing.

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand customer behavior and review patterns. Multiple visualizations were created to study rating distributions, sentiment balance, review lengths, and restaurant popularity.

1. Most customer ratings are skewed towards higher values, indicating general satisfaction.
2. Negative reviews tend to be longer, suggesting customers provide detailed feedback when dissatisfied.
3. Food and service are the most frequently discussed aspects in reviews.
4. High review volume does not always correlate with higher average ratings.

6. Hypothesis Testing & Statistical Validation

To validate insights derived from EDA, formal hypothesis testing was performed. Independent two-sample t-tests were used to compare review lengths and ratings across sentiment groups. A chi-square test of independence was applied to examine the relationship between food-related mentions and sentiment polarity. All tests resulted in statistically significant outcomes, confirming that the observed patterns were not due to random chance.

7. Text Preprocessing & Feature Engineering

Text preprocessing is a critical step in NLP projects. The review text was converted to lowercase, punctuation and special characters were removed, and stopwords were filtered out. Stemming was applied using the Porter Stemmer to reduce words to their root forms. In addition to textual features, numerical features such as review length and aspect-based indicators were engineered to enhance model performance.

8. Text Vectorization

TF-IDF (Term Frequency–Inverse Document Frequency) vectorization was used to convert text into numerical form. This technique assigns higher importance to words that are frequent in a review but rare across the dataset, making it effective for sentiment classification tasks.

9. Machine Learning Model Development

Two supervised learning models were implemented: Logistic Regression and Multinomial Naive Bayes. Logistic Regression was selected as the final model due to its superior performance and interpretability. Class imbalance was handled using balanced class weights, and hyperparameter tuning was performed using GridSearchCV.

10. Model Evaluation & Business Impact

Model performance was evaluated using accuracy, precision, recall, and F1-score. A confusion matrix was used to analyze misclassification patterns, particularly false negatives, which represent dissatisfied customers that may otherwise be overlooked. From a business perspective, minimizing false negatives is critical for proactive issue resolution.

11. Model Interpretability & Insights

Model interpretability was achieved by analyzing feature coefficients from the Logistic Regression model. This analysis revealed keywords strongly associated with positive and negative sentiment. Additionally, restaurant-level sentiment aggregation enables management to identify top-performing and underperforming restaurants for targeted improvement initiatives.

12. Deployment Readiness

The final trained model and TF-IDF vectorizer were saved using pickle format. This ensures the model can be easily integrated into real-world applications such as dashboards, APIs, or customer feedback monitoring systems.

13. Conclusion

This project demonstrates a complete end-to-end data science workflow, combining exploratory analysis, statistical validation, NLP preprocessing, machine learning modeling, and deployment readiness. The solution provides both predictive capability and actionable business insights, making it valuable for restaurant owners and food delivery platforms alike.