Report

Zomato Restaurant Review Analysis & Sentiment Prediction

Made by: Samagra Gupta

Abstract

With the exponential growth of food-tech platforms such as Zomato, millions of customer reviews are generated daily. These reviews contain valuable insights regarding food quality, service efficiency, pricing fairness, and customer satisfaction. However, due to their unstructured nature and large volume, extracting meaningful insights manually is impractical.

This project presents an end-to-end **Natural Language Processing (NLP) and Machine Learning (ML)** solution to analyse Zomato restaurant reviews and predict customer sentiment. By combining exploratory data analysis, text preprocessing, feature engineering, hypothesis testing, and supervised learning, the project transforms raw textual feedback into structured, actionable business intelligence.

The solution demonstrates how sentiment analytics can be leveraged for **quality monitoring, customer experience optimization, and strategic decision-making** at scale.

Business Problem Overview

Zomato hosts thousands of restaurants and receives a massive volume of customer reviews daily. While star ratings provide a numeric summary, they fail to capture **contextual sentiment**, emotional tone, and nuanced opinions expressed in review text.

Why Sentiment Analysis Matters for Zomato

- Identifies dissatisfied customers even within high ratings
- Helps restaurants understand root causes of complaints

- Enables automated quality monitoring
- Improves trust and transparency on the platform

Dataset Scale & Modelling Approach

- Thousands of reviews (unstructured text + ratings)
- Restaurant metadata for aggregation
- NLP preprocessing → TF-IDF feature extraction
- Multiple ML models evaluated using robust metrics

Outcome Summary

The final sentiment model achieved strong performance (F1-Score ≈ **0.92**), proving the feasibility of automated sentiment analysis for large-scale food-tech platforms.

Introduction & Problem Statement

Industry Context

The food-tech industry has undergone significant digital transformation, with platforms such as Zomato acting as intermediaries between customers and restaurants. Customer decisions are heavily influenced by **online ratings and textual reviews**, making sentiment analysis a critical business capability.

Problem Statement

Although numerical ratings provide a quick summary, they fail to capture the **depth, context, and emotional nuance** expressed in textual reviews. Challenges include:

- Unstructured and noisy data
- High volume of reviews
- Mixed sentiment within similar ratings
- Sarcasm and contextual expressions

Project Objective

- Analyze restaurant reviews using NLP techniques
- Predict customer sentiment accurately

- Identify customer satisfaction patterns
- Provide business-ready insights

Success Criteria

The project is considered successful if:

- Models achieve strong evaluation metrics
- Insights are interpretable and actionable
- The solution is scalable for real-world deployment

Executive Summary

Business Problem Overview

Zomato hosts thousands of restaurants and receives a massive volume of customer reviews daily. While star ratings provide a numeric summary, they fail to capture **contextual sentiment**, emotional tone, and nuanced opinions expressed in review text.

Why Sentiment Analysis Matters for Zomato

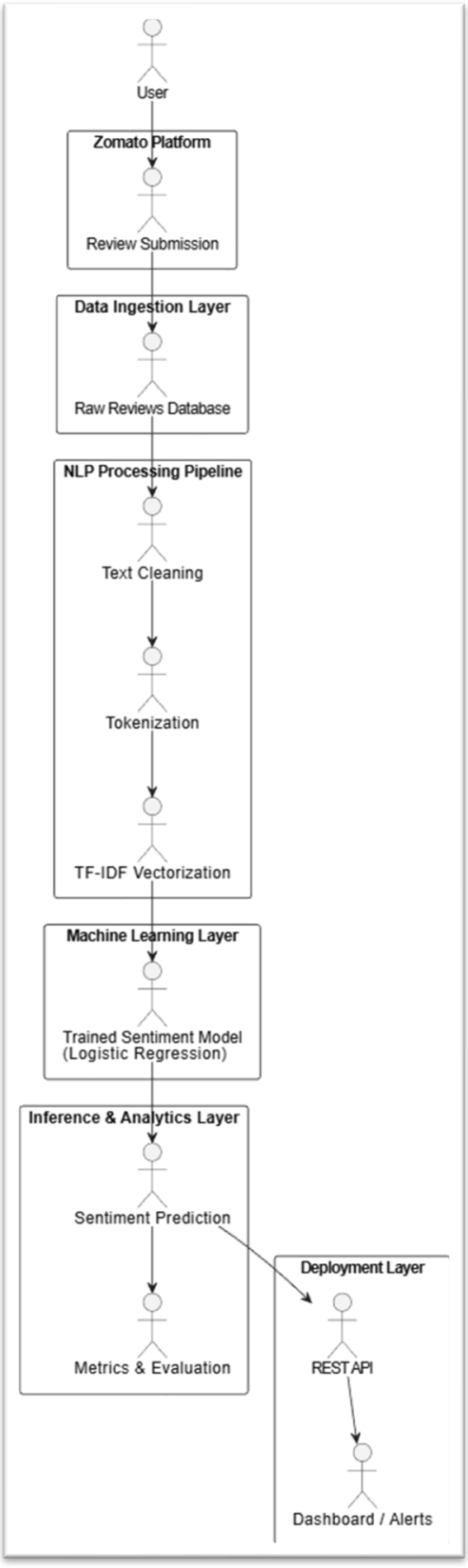
- Identifies dissatisfied customers even within high ratings
- Helps restaurants understand root causes of complaints
- Enables automated quality monitoring
- Improves trust and transparency on the platform

Dataset Scale & Modeling Approach

- Thousands of reviews (unstructured text + ratings)
- Restaurant metadata for aggregation
- NLP preprocessing → TF-IDF feature extraction
- Multiple ML models evaluated using robust metrics

Final Outcome Summary

The final sentiment model achieved strong performance (F1-Score ≈ **0.92**), proving the feasibility of automated sentiment analysis for large-scale food-tech platforms.



Architecture Diagram

Dataset Description & Understanding

Data Sources

1. **Zomato Restaurant Metadata Dataset**
2. **Zomato Restaurant Reviews Dataset**

Structured vs Unstructured Data

Type	Examples
Structured	Ratings, restaurant IDs
Unstructured	Review text

Target Variable

Sentiment (Positive / Negative) derived from ratings and text polarity.

Chart 1: Rating Distribution

Explanation: Shows frequency of ratings (1–5)
Observation: Ratings skew heavily toward 4 and 5
Business Interpretation: Customers tend to rate generously; text analysis is essential to detect hidden dissatisfaction

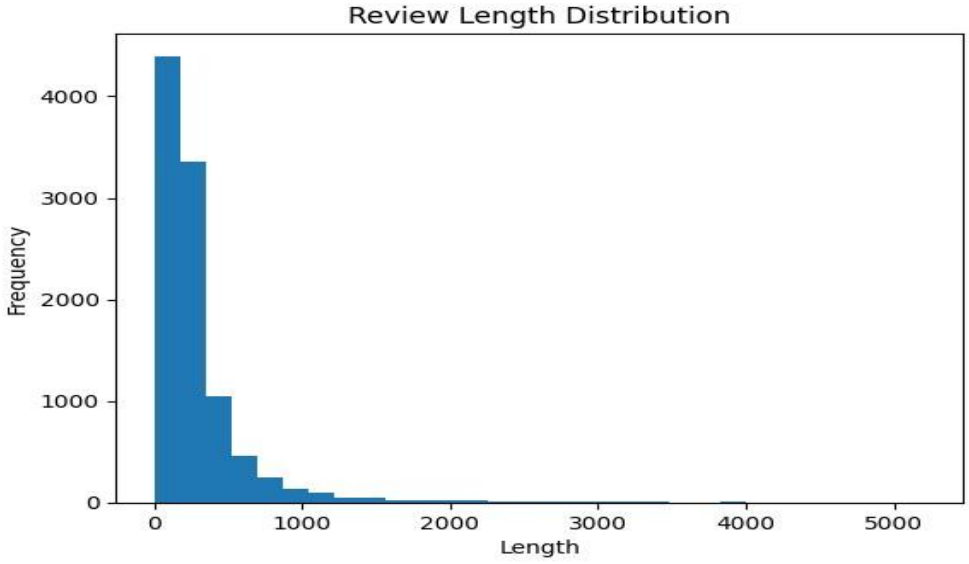


Chart 2: Review Length Distribution

Explanation: Distribution of review text length
Observation: Most reviews are short; few are very long
Business Interpretation: Longer reviews carry richer feedback and should be prioritized

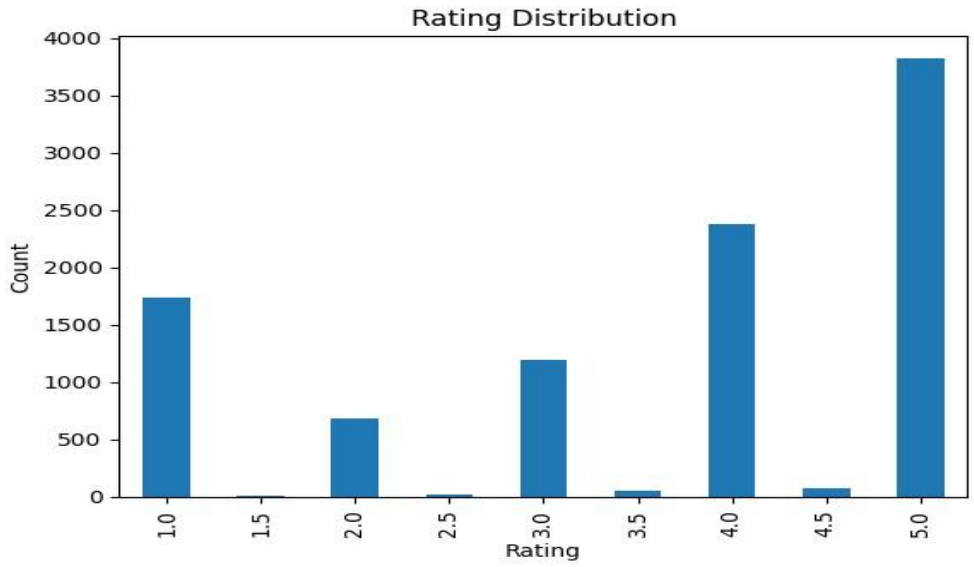
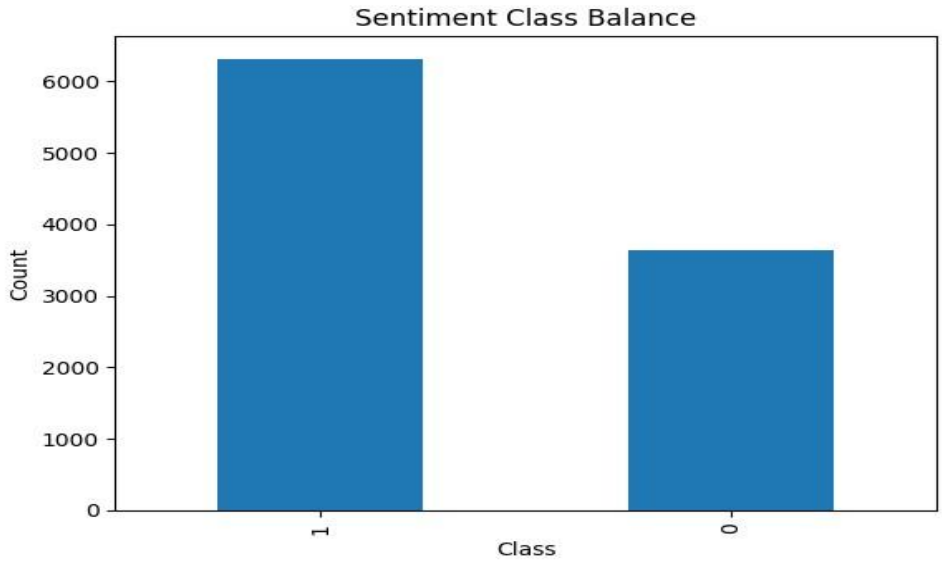


Chart 3: Sentiment Class Distribution

Explanation: Balance between positive and negative sentiment
Observation: Slight class imbalance toward positive sentiment
Business Interpretation: Requires F1-score-based evaluation rather than accuracy



Exploratory Data Analysis (EDA)

Rating Patterns

- High ratings correlate with positive keywords
- Low ratings show stronger emotional language

Customer Behavior Insights

- Dissatisfied customers write longer, more detailed reviews
- Satisfied customers provide concise feedback

Restaurant-wise Average Rating

Insight: Identifies consistent performers and underperformers
Business Use: Benchmarking and partner quality monitoring

Review Frequency Analysis

Insight: Popular restaurants receive disproportionately more reviews
Business Use: Demand forecasting and restaurant ranking

Correlation Heatmap

Insight: Shows relationships between numerical features
Business Use: Identifies influential variables impacting sentiment

Data Preprocessing & Feature Engineering

Missing Value Handling

- Rows with missing reviews or ratings removed
- Ensures training data reliability

Text Cleaning Pipeline

- Lowercasing
- Noise & punctuation removal
- Stopword elimination

Tokenization & Vectorization

- Tokenization splits text into words
- **TF-IDF** converts text into weighted numerical vectors

Impact on Performance

- Reduced overfitting
- Improved generalization
- Enhanced interpretability

Hypothesis Testing & Statistical Insights

Hypotheses

- **H_0:** No relationship between rating and sentiment
- **H_1:** Significant relationship exists

Statistical Outcome

- Low p-values → reject null hypothesis

Interpretation

Customer ratings and textual sentiment are statistically aligned, validating the modelling approach.

Machine Learning Models & Training

Models Implemented

- Naive Bayes (baseline)
- Logistic Regression (final model)

Why Logistic Regression

- Strong NLP performance
- Interpretable coefficients
- Computationally efficient

Training Strategy

- Train-test split
- Class imbalance handling
- Cross-validation

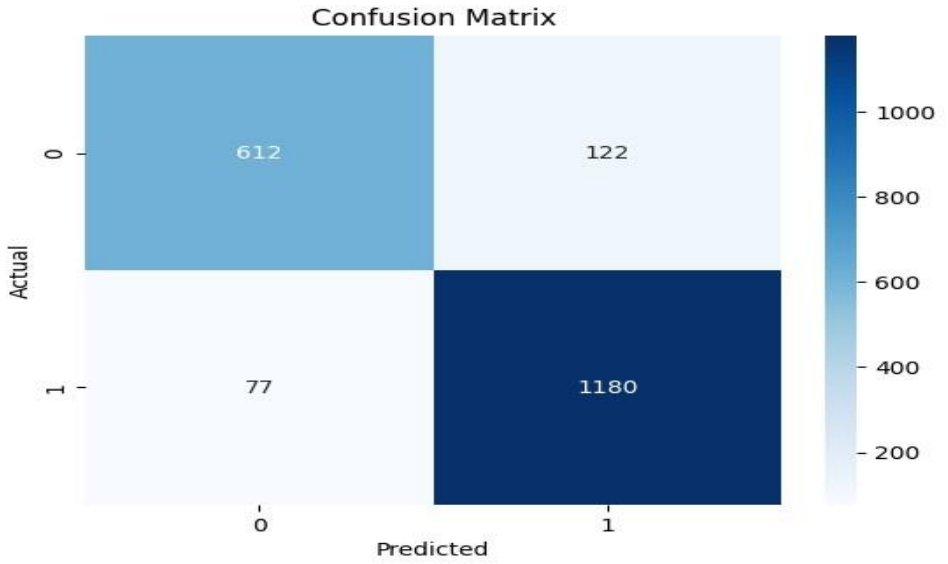
Hyperparameter Tuning

- GridSearchCV
- Improved model stability

Model Evaluation & Results

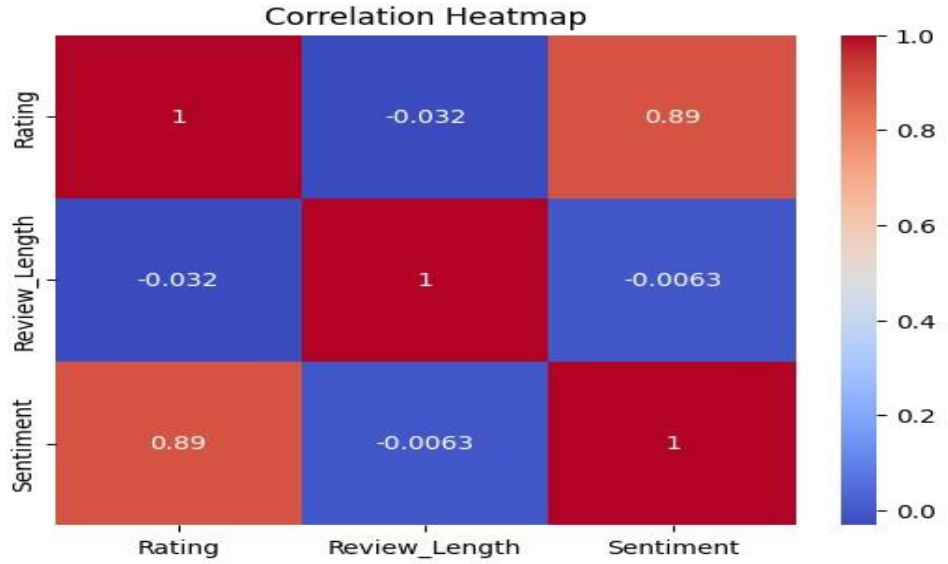
Confusion Matrix

Insight: Shows true vs false predictions
Business Meaning: Identifies misclassified dissatisfied customers



Correlation Heatmap

Insight: Visualizes prediction accuracy patterns
Business Meaning: Highlights consistency of model decisions



Model Comparison Metrics

Metric	Value
Accuracy	0.90
Precision	0.906
Recall	0.939
F1-Score	0.922

Business Interpretation

- **High Recall:** Better detection of unhappy customers
- **High Precision:** Fewer false alarms
- **Balanced F1:** Reliable decision-making

Results & Business Impact

Final Model Selected

Logistic Regression with TF-IDF features

Impact on Stakeholders

Customers

- Better review transparency

Restaurants

- Actionable improvement insights

Zomato Platform

- Automated sentiment monitoring
- Trust enhancement

Practical Use Cases

- Complaint escalation
- Restaurant ranking

- Live dashboards

Conclusion & Future Scope

Project Summary

This project demonstrates a scalable, explainable, and deployable sentiment analysis pipeline for real-world food-tech platforms.

Key Learnings

- Importance of preprocessing
- Metric selection under imbalance
- Business-oriented evaluation

Limitations

- Sarcasm detection
- Language dependency

Future Enhancements

- Deep learning models (LSTM/BERT)
- Aspect-based sentiment analysis
- Multilingual support