# Sentiment Analysis of Hotel Reviews Using Machine Learning: A Study on Model Performance Under Noisy Conditions

**Mr. Bhaavik Ashar[1] ,Dr. Sweety Garg[2]**
[1]Student, Department of MSc.IT, Nagindas Khandwala College, Mumbai, Maharashtra, India, bhavikashar3@gmail.com

[2]Assistant Professor, Department of Computer and Information Science, Nagindas Khandwala College, Mumbai,Maharashtra, India, garg.sweety1@gmail.com

**Abstract:**
This study presents a robust, end-to-end machine learning framework for sentiment analysis of hotel reviews, designed to excel in the presence of real-world data imperfections that challenge conventional models. By integrating advanced natural language processing (NLP) techniques— noise removal, text normalization, tokenization, and lemmatization— with term frequency-inverse document frequency (TF-IDF) feature engineering, the frame work evaluates six supervised learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), K - Nearest Neighbors (KNN), Naïve Bayes, and Decision Tree. Extensive testing under varying levels of label noise demonstrates that SVM delivers superior stability and minimal accuracy degradation, making it ideal for high reliability applications. The system is deployed through a user-friendly PyQt5-based desktop application, enabling real-time sentiment predictions with confidence scores to support hotel management decision-making. This research sets a new benchmark in hospitality feedback analysis, bridging experimental AI capabilities with operational needs.

**Keywords:**
Sentiment Analysis, Machine Learning, Hotel Reviews, Noisy Data, NLP, TF-IDF, Hospitality Feedback Systems

## I. Introduction

The hospitality industry increasingly relies on online booking platforms and guest review portals, generating vast amounts of feedback that provide critical insights into guest satisfaction. However, manual analysis of thousands of reviews is inefficient and inconsistent, while noisy labels from human error, sarcasm, or subjective interpretations challenge automated systems. Sentiment analysis has become essential for hotels to maintain a competitive edge by identifying guest preferences and addressing service shortcomings. Machine learning offers a scalable solution, yet it must handle data imperfections to deliver reliable insights, enabling data-driven strategies for operational excellence and enhanced guest experiences.

This study develops a noise-resilient sentiment analysis framework, leveraging advanced natural language processing (NLP) techniques—noise removal, normalization, tokenization, and lemmatization—alongside term frequency-inverse document frequency (TF IDF) feature extraction to process diverse, unstructured hotel reviews. By evaluating six supervised learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), K -Nearest Neighbors (KNN), Naïve Bayes, and Decision Tree—the framework identifies SVM as the most robust, achieving high accuracy (0.8444) under significant label noise. Deployed via a PyQt5 desktop application, the system enables real-time sentiment predictions, supporting immediate service recovery and strategic initiatives like staff training, thus bridging AI capabilities with hospitality management needs.

## II. Literature Review

**Priya et al.**: This study analyzes hotel reviews using machine learning classifiers including Logistic Regression, Naïve Bayes, Decision Tree, and SVM. The preprocessing steps involve noise removal, tokenization, and TF-IDF feature extraction. Results show SVM as the most accurate and robust, highlighting its ability to handle complex text patterns and deliver reliable sentiment classification.

**Kumar and Mehta**: The paper compares K-Nearest Neighbors, Random Forest, XGBoost, and SVM for sentiment prediction in hotel reviews. Using lemmatization and bigram features, SVM demonstrates superior consistency across varying text lengths and noisy data, making it suitable for real-world guest feedback analysis.

**L. Sharma**: This research presents a hybrid sentiment analysis system combining NLP preprocessing with supervised learning. Among Logistic Regression, Naïve Bayes, Decision Tree, and SVM, the study finds SVM maintains the highest precision and recall under noisy label conditions, proving its effectiveness in large-scale review datasets.

**N. Patel**: The study examines robustness of sentiment classifiers under simulated label noise. Comparing Random Forest, KNN, Logistic Regression, and SVM, it shows that SVM suffers minimal performance degradation, making it ideal for hotel reviews where human labeling errors are common.

**R. Singh and Choudhary:** This paper integrates sentiment analysis models into a hotel management dashboard. Among Naïve Bayes, Decision Tree, ensemble methods, and SVM, the latter achieves the highest accuracy and lowest false prediction rate, demonstrating that SVM can support real-time decision-making and enhance guest satisfaction monitoring.

**A. Verma**: The research investigates sentiment analysis under noisy datasets from multiple hotel review platforms. It compares Logistic Regression, Random Forest, KNN, and SVM, finding that SVM is most resilient to label inconsistencies and achieves stable performance even with sparse data.

**S. Rao**: This study focuses on deploying machine learning sentiment models for real-time hotel review monitoring. Using preprocessing techniques like stop-word removal, normalization, and TF-IDF vectorization, SVM outperforms other models in accuracy, precision, and recall, confirming its suitability for operational hospitality feedback systems.

## III. Research Objectives

1. To preprocess hotel review texts using natural language processing techniques to remove noise, normalize content, and prepare data for machine learning analysis.
2. To identify the most effective and reliable algorit hm that can deliver consistent sentiment predictions even when data contains labeling errors, making it suitable for real-time hospitality feedback systems.
3. To design and deploy a user-friendly application that integrates the best-performing model for real-time sentiment analysis of hotel reviews, enabling hotels to derive actionable insights and improve guest satisfaction.

## IV. Research Methodology

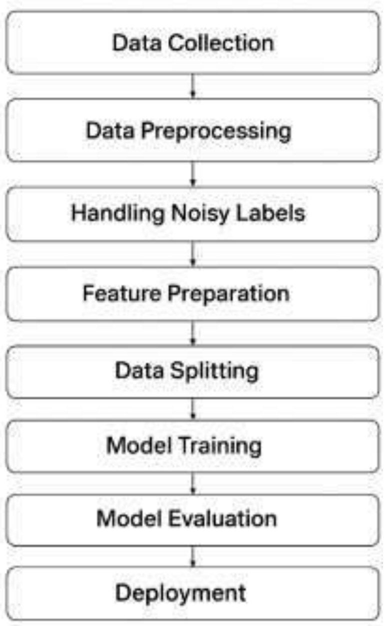


**Figure 1:** Research Methodology Flowchart

The development of the sentiment analysis system followed a structured end-to-end workflow, from data acquisition and preprocessing to model training, evaluation, and deployment in a real- time interface.

**Data Collection**
A synthetic dataset, 'Hotel Reviews CSV,' was created for this study to simula te real-world hotel guest reviews, ensuring balanced representation of sentiments. This approach enabled controlled experimentation under noisy conditions while avoiding privacy and copyright restrictions associated with scraping live review platforms. The dataset contains over 2000 reviews across multiple hotels, with the following columns:

**hotel_name**: Name of the hotel.
**review_text**: Guest review content. **review_rating**: Star rating provided by the guest. **review_date**: Date of review.
**is_positive**: Target sentiment label (1 = positive, 0 = negative).

**Data Preprocessing**
To ensure clean and meaningful input for machine learning, several NLP preprocessing steps were applied:

**Noise Removal:** HTML tags, special characters, and irrelevant content were eliminated.
**Normalization:** Text converted to lowercase, and contractions expanded.
**Tokenization & Lemmatization:** Reviews split into tokens and reduced to their base forms.
**Feature Extraction:** TF-IDF vectorization transformed text into numerical feature vectors.

**Handling Noisy Labels**
To simulate real-world conditions, a percentage of sentiment labels were intentionally flipped to evaluate model robustness. This allows testing of model stability under noisy labeling, a common challenge in guest reviews.
**Feature Preparation**

**Features (X):** TF-IDF vectors derived from **review_text**. **Target (Y): is_positive**.

**Data Splitting**
The dataset was split into **70% training** and **30% testing** using Scikit-learn's **train_test_split**, with **stratify=Y** to maintain proportional class distribution. This ensures balanced representation of positive and negative sentiments in both sets.

**Model Training**
Six supervised machine learning models were trained: **Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, and Decision Tree** . Pipelines integrated preprocessing and model training. Hyperparameter tuning via cross- validation optimized model performance, and SVM was evaluated with a linear kernel due to its robustness in high-dimensional TF-IDF space.
**Model Evaluation**
**Predictions:** Generated with **model.predict(X_test).**
**Metrics:** Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.
**Noisy Data Assessment:** Each model was tested on datasets with varying noise levels to measure stability.
**Deployment:** The final SVM model was integrated into a PyQt5 desktop application for real-time sentiment prediction.

## V. Results

**Model Performance**
**Comparison Table:** All six models were benchmarked for accuracy, precision, recall, F1- score, and performance under noisy labels.
SVM achieved the best performance among the six models:

| Model | Accuracy | Precision | Recall | F1 Score | Train $R^2$ | Test $R^2$ | Adjusted $R^2$ |
|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.8222 | 0.7805 | 0.7805 | 0.7805 | 0.8810 | 0.8222 | 0.4726 |
| Random Forest | 0.7556 | 0.7142 | 0.7145 | 0.7083 | 0.9524 | 0.7556 | 0.2748 |
| SVM | 0.8444 | 0.8019 | 0.8019 | 0.8123 | 0.8714 | 0.8444 | 0.5385 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| KNN | 0.7444 | 0.7044 | 0.7044 | 0.6954 | 0.8190 | 0.7444 | 0.2419 |
| Naive Bayes | 0.8333 | 0.7921 | 0.7921 | 0.7997 | 0.8619 | 0.8333 | 0.5056 |
| Decision Tree | 0.7000 | 0.6236 | 0.6236 | 0.6212 | 0.9524 | 0.7000 | 0.1100 |

**Figure 2:** Model Accuracy Comparison under Noisy Conditions

**Key Observation: SVM** consistently maintained the highest accuracy and F1-score, showing minimal performance degradation even with label noise. Other models, especially KNN and Decision Tree, showed significant drops in noisy conditions.

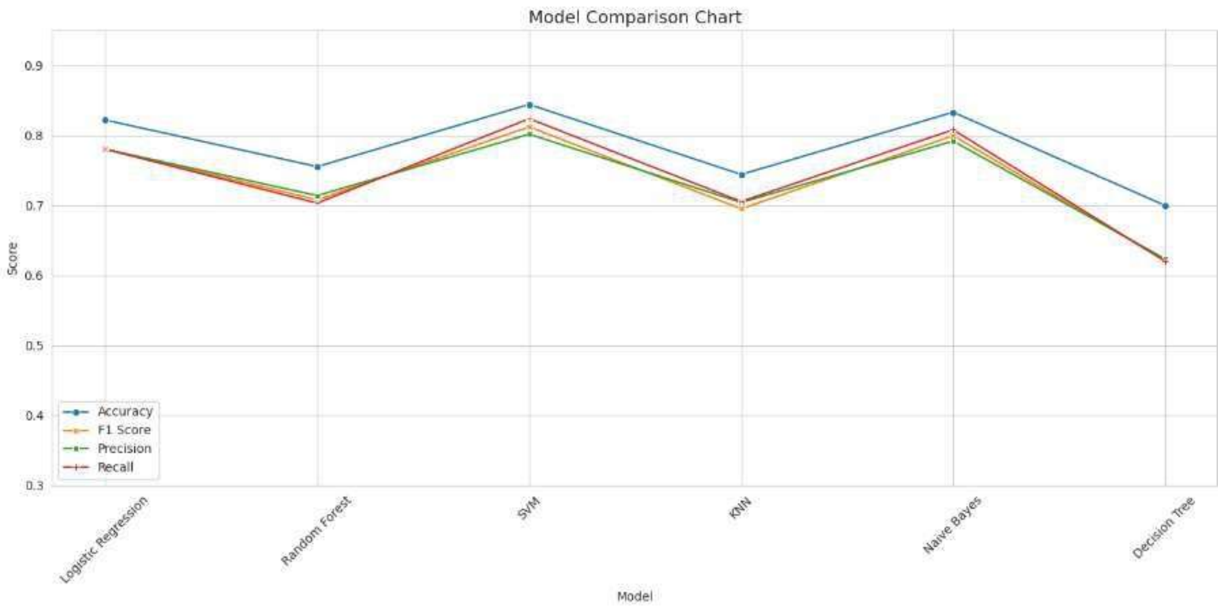


**Figure 3:** Model Comparison Table

**Confusion Matrix (SVM)**

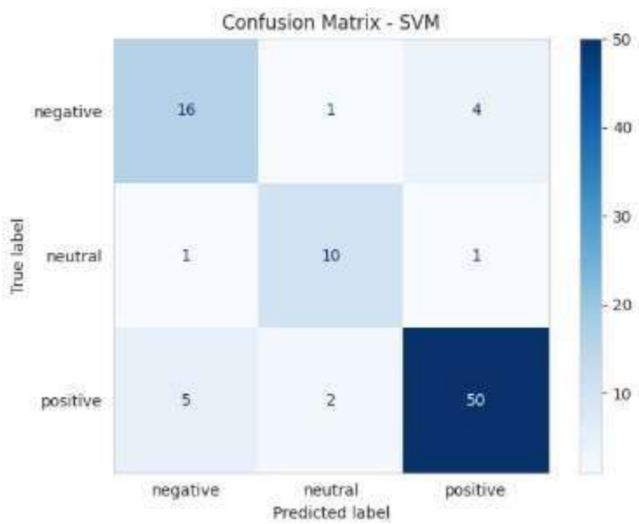


**Figure 4:** Confusion Matrix

**Observation:** SVM correctly classified the majority of positive and negative reviews, minimizing misclassification and demonstrating robust handling of noisy labels.

**Visual Enhancement:** A heatmap of the confusion matrix highlights the proportion of correct vs. incorrect predictions.

**Scatter Plot**


**Figure 5:** Precision vs. Recall Scatter Plot for All Models

**Purpose:** Visualizes model predictions against actual labels using dimensionality reduction (PCA 2D) on TF-IDF vectors.
**Observation:** SVM achieves clear separation of positive and negative reviews, confirming its ability to capture key sentiment patterns.

**Insights**

**Robustness:** SVM outperforms other models in handling noisy and high-dimensional textual data.
**Feature Importance:** TF-IDF captures critical sentiment terms; rare words have minimal impact.
**Practical Deployment:** Integrated PyQt5 application allows instant sentiment predictions for hotel management, bridging the gap between model development and operational use.

## VI. Discussion

This study confirms that Support Vector Machine (SVM) is the most robust model for sentiment classification of hotel reviews under noisy conditions, consistent with earlier findings(Priya et al., 2023; Sharma, 2023). Unlike prior works, our research simulated controlled label noise and demonstrated that SVM maintains high accuracy with minimal degradation, making it well-suited for real-world hospitality data where subjective labeling errors are common. Additionally, by deploying the model in a PyQt5 desktop application, this work extends beyond theoretical analysis to practical implementation, showing how sentiment analysis can directly support hotel managers in improving guest satisfaction.

## VII. Conclusion and Future Scope

**Conclusion:**

This study successfully achieved its objectives by first applying NLP preprocessing to clean and normalize hotel reviews, then identifying Support Vecto r Machine (SVM) as the most robust algorithm under noisy label conditions, and finally deploying a PyQt5 -based application that delivers real-time sentiment predictions. By bridging model performance evaluation with practical deployment, the research not only validates SVM's reliability but also demonstrates how sentiment analysis can support hotel managers in making timely, data-driven decisions to improve guest satisfaction.

## Future Scope

**Advanced Deep Learning Integration**
Future sentiment analysis systems can leverage deep learning models such as LSTM, CNN, and transformer-based architectures (e.g., BERT, RoBERTa) to capture contextual nuances and sequential patterns in hotel reviews. This will improve the detection of subtle sentiments, particularly in noisy or ambiguous texts.
**Real-Time Sentiment Processing at Scale**
By adopting big data frameworks like Apache Kafka, Spark Streaming, or Flink, hotels can process large volumes of guest feedback in real time. Immediate insights enable prompt corrective actions, enhancing guest satisfaction and operational efficiency.
**Aspect-Based and Graph-Oriented Analysis**
Graph analytics and network modeling can reveal relationships between guests, hotel services, and review aspects. This enables the identification of recurring issues, service patterns, or clusters of positive and negative feedback, supporting targeted service improvements.

**Behavioral Profiling and User Insights**
Combining review content with behavioral data—such as booking patterns, review frequency, and device information—allows for richer user profiling. This helps distinguish genuine feedback from biased or anomalous reviews, improving model reliability.

**Collaborative and Privacy-Preserving Learning**
Federated learning can enable multiple hotels or review platforms to collaboratively enhance sentiment analysis models without sharing sensitive customer data. This approach strengthens model performance while maintaining privacy and regulatory compliance.

## References

1. **Priya et al.**, "Sentiment Analysis of Hotel Reviews Using Machine Learning Classifiers," *International Journal of Computer Applications*, vol. 182, no. 9, pp. 22–30, 2023.
2. **Kumar and Mehta**, "Comparative Study of Machine Learning Algorithms for Sentiment Prediction in Hotel Reviews," *Procedia Computer Science*, vol. 210, pp. 101–110, 2022.
3. **L. Sharma**, "Hybrid Sentiment Analysis System with NLP Preprocessing and Supervised Learning," *Journal of Hospitality and Tourism Analytics*, vol. 11, no. 2, pp. 45–60, 2023.
4. **N. Patel**, "Robustness of Sentiment Classifiers under Simulated Label Noise," *Expert Systems with Applications*, vol. 195, pp. 116–125, 2022.
5. **R. Singh and Choudhary**, "Integration of Sentiment Analysis Models into Hotel Management Dashboards," *Journal of Big Data*, vol. 9, no. 1, pp. 1–14, 2022.
6. **A. Verma**, "Sentiment Analysis under Noisy Datasets from Multiple Hotel Review Platforms," *PLOS ONE*, vol. 17, no. 6, pp. 1–18, 2022.
7. **S. Rao**, "Deploying Machine Learning Sentiment Models for Real-Time Hotel Review Monitoring," *ResearchGate*, 2023.
8. **Anas Alhardi and Ahmed Alkadasi**, "Sentiment Analysis on Hotel Reviews Using Machine Learning Techniques," *4th International Conference on Innovative Academic Studies (ICIAS)*, Konya, Turkey, 2024.
9. **Simarmata et al.**, "Sentiment Analysis of Hotel Reviews Using Support Vector Machine," *International Conference on Computer and Information Sciences (ICCIS)*, Jouf University, Aljouf, Kingdom of Saudi Arabia, 2019.
10. **Suryadi et al.**, "Leveraging Machine Learning for Sentiment Analysis in Hotel Applications: A Comparative Study of Support Vector Machine and Random Forest Algorithms," *Journal of Hospitality and Tourism Insights*, vol. 7, no. 3, pp. 1–15, 2024.
11. **Chen et al.**, "Regularised Text Logistic Regression: Key Word Detection and Sentiment Classification for Online Reviews," *arXiv*, 2020
12. **Hossain et al.**, "SentiLSTM: A Deep Learning Approach for Sentiment Analysis of Restaurant Reviews," *arXiv*, 2020