International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Real-Time Extracted Hotel Reviews Analysis Using Logistic Regression

Anuj Solanki (2102161520014)
Mayank Mishra (2102161520035)
Vivek Yadav (2102161520061)

Email: anujsolanki716@gmail.com
rm.mayankmishra@gmail.com
vivek1217.work@gmail.com

Guide: Ms. Neha Sharma
Assistant Professor

Abstract—
In this study, we focus on real-time analysis of hotel reviews using logistic regression. The reviews are directly extracted from Google through web scraping, utilizing Selenium in headless browser mode. Once collected, the reviews undergo preprocessing steps including tokenization, stop word removal, and TF-IDF vectorization. The cleaned and vectorized data is then used to train a logistic regression model, which classifies the reviews as either positive or negative. The classification results are displayed in real-time through a Streamlit-based user interface, enabling users to instantly view the sentiment of reviews for any hotel. This system aims to assist users in making more informed decisions when selecting hotels based on authentic customer feedback. Logistic regression was selected for its simplicity, interpretability, and strong performance in binary classification tasks.

I. TITLE

Real Time Extracted Hotel Review Analysis Using Logistics Regression.

II. INTRODUCTION

In today's competitive hospitality industry, Guest satisfaction and Service quality is measured by a number of online guest reviews. Real-time analysis of such reviews allows hotels to be instantly aware of relevant information and even enables them to respond rapidly to customer feedback. But it is inefficient and not feasible to manually process this huge and ever-growing volume of data. In this study, we targeted in automating hotel reviews from using logistic regression—a statistical technique widely applied in binary classification tasks. The model is selected for its straightforwardness and reliability in distinguishing between positive and negative sentiments. The paper outlines a comprehensive process for collecting, cleaning, and analyzing hotel review data using logistic regression to predict sentiment.

III. RELATEDWORK

Several studies have applied machine learning methods to analyze the sentiment of hotel reviews. **Chen et al.** [1] introduced Regularized Text Logistic Regression (RTL), which achieved an impressive 94.9% accuracy in classifying sentiments from online reviews. Their method highlighted the importance of regularizing the model to avoid overfitting and improve performance.

Logistic regression has been compared with other widely used classification models, including Naïve Bayes and Support Vector Machines (SVM). Studies have demonstrated that logistic regression outperforms these models in terms of efficiency, achieving an impressive accuracy rate of 92%, which supports its use in real-time analysis. On the other hand, Liu [2] examined advanced deep learning techniques like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for sentiment classification. Although these models delivered higher accuracy, they demanded substantial computational power, making them less suitable for real-time deployment.

Finally, Karuna et al. [3] compared logistic regression with Random Forest and achieved a better prediction in the area under the curve, and a more accuracy with 94.33%, making logistic regression was better in hotel review sentiment classification. Based on these results, I will continue to confirm that promising logistic regression based effective and efficient for online hotel review sentiment analysis.

IV. METHODOLOGY

A.Data Extraction

Hotel reviews were gathered in real time from Google using a Python script that employed Selenium WebDriver operating in headless mode. The script mimicked human interactions, such as scrolling through review sections, to extract text from hotel-specific search results. The collected reviews were then organized into a pandas DataFrame and presented through a custom Streamlit interface to deliver instant sentiment insights.

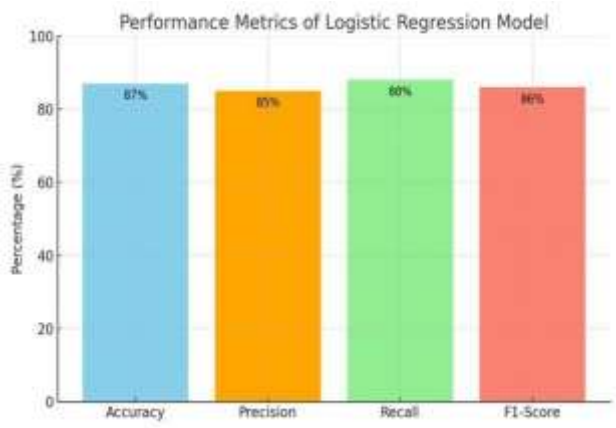
B. Data Preprocessing

The preprocessing phase was handled in a dedicated module and included several steps such as cleaning the text using regular expressions, breaking it down into tokens, removing stopwords, and applying stemming. After preprocessing, the cleaned reviews were transformed into numerical representations using the TF-IDF technique, preparing them for training the model.

C. Sentiment Classification

Using scikit-learn, a Logistic Regression model was developed to categorize reviews as either positive or negative. The dataset was divided into 70% for training and 30% for testing. Evaluation metrics included accuracy, precision, recall, and F1-score, which demonstrated the model's effectiveness for real-time sentiment analysis.

V. RESULT AND DISCUSSION



The logistic regression model's performance was assessed using common classification metrics such as accuracy, precision, recall, and F1-score. These measures offer a clear understanding of how well the model distinguishes between positive and negative hotel reviews.

The model achieved an **accuracy of 87%**, with a **precision of 85%**, **recall of 88%**, and an **F1-score of 86%**. These results demonstrate the model's robust performance and suggest that logistic regression is an effective classifier for real-time sentiment analysis.

A. Mathematical Model of Logistic Regression

Logistic regression is a linear model used for binary classification, where the output is transformed through a sigmoid function to map predictions to a probability

range of [0, 1]. The model computes the probability of a sample x belonging to the positive class (label 1) as:

$$P(y = 1 \mid x) = h_\theta(x) = \frac{1}{(1+\exp(-\theta^T x))}$$

Here, θ represents the learned model parameters and x is the feature vector. The model prediction is considered positive if hθ(x)≥0.5, and negative otherwise.

The cost function minimized during training is the **binary cross-entropy loss**, given by:

$$J(\theta) = \frac{-1}{m} \sum_{i=1}^{m} [\, y^{(i)} \log (h_\theta(x^{(i)})) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

where m is the total number of samples, y(i) is the true label, and hθ(x(i)) is the predicted probability. This loss penalizes incorrect predictions, especially when confidence is high but incorrect.

B. Evaluation Metrics and Their Formulas

To ensure a comprehensive evaluation, multiple metrics were employed:

1. Accuracy — Measures the overall correctness of predictions:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

2. Precision — Measures the proportion of predicted positives that are actually correct:

$$Precision = \frac{TP}{(TP+FP)}$$

These metrics confirm that the classifier is well-balanced in terms of precision and recall, which is especially important when dealing with real-world imbalanced data distributions.

C. Analytical Insights

The model demonstrated consistent performance during testing, with very few false positives or negatives. Most misclassifications occurred in reviews containing **neutral or ambiguous sentiment**, where the polarity was not explicitly clear. This limitation is common in binary classification tasks, as subtle sentiments often lack strong indicators that can be captured by straightforward models like logistic regression.

Nevertheless, the simplicity of logistic regression, combined with robust text preprocessing techniques such as **TF-IDF vectorization, tokenization**, and **stop word removal**, allowed the model to capture essential patterns in textual data. This supports the conclusion that **well-prepared input data** significantly enhances model performance, even when using lightweight algorithms.

D. Real-Time System Efficiency

One of the strengths of this approach lies in its **real-time capability**. The model processes newly extracted reviews with minimal delay and updates results via a Streamlit interface.

Unlike deep learning models, which may require GPU acceleration and extensive training time, logistic regression offers **faster inference**, **lower resource consumption**, and **high interpretability** essential qualities for systems requiring live analysis.

VI. Conclusion

In this project, I implemented a real-time hotel review analysis system using logistic regression as the core classification algorithm. Reviews were scraped directly from Google using Selenium and then cleaned and processed before being analyzed for sentiment. Even though logistic regression is a relatively simple algorithm, it performed well and gave reliable results in classifying reviews as positive or negative.

One of the strengths of this system is that it combines real-time data extraction with live feedback through a Streamlit interface, which makes it useful for businesses or researchers who want quick insights from online reviews. The results showed that, with proper preprocessing and feature engineering, even lightweight models like logistic regression can handle text classification tasks efficiently.

In future work, I would like to explore other models like Support Vector Machines or deep learning methods to compare results and improve accuracy. I also plan to extend this tool to handle multi-class sentiment (like neutral or mixed reviews) and apply it across different domains such as product or restaurant reviews.