International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

ENHANCING RENTAL LISTINGS USING MACHINE LEARNING AND NLP

Dr. S. Venkata Achuta Rao[1], N. Nikhila[2], D. Yashwanth[2], G. Deepak[2], Y.V.V. Vardhan kumar[2]
[1]Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (AI&ML)
[1,2]Sree Dattha Institute of Engineering and Science, Ibrahimpatnam, 501510, Telangana

Received: 09-07-2025 Accepted: 23-08-2025 Published: 30-08-2025

ABSTARCT

India's rental market has grown rapidly due to urbanization and population increases, particularly in cities like Mumbai, Delhi, and Bengaluru, where rental demand has surged by 20–30%. Despite this growth, most property listings lack data-driven insights into customer preferences. Traditionally, landlords relied on newspaper ads, brokers, and intuition to market properties and set prices, with little personalization or analytical feedback. This approach made it difficult to accurately gauge tenant interest and often resulted in inefficient, generic marketing. The proposed system addresses these limitations by integrating Natural Language Processing (NLP) and machine learning to analyze sentiments in property listings and viewer feedback. By processing descriptions and extracting key renter-attracting features, the system uses predictive models to classify sentiment from reviews, inquiries, and social media. This enables landlords to optimize pricing, tailor listings, and improve marketing strategies based on viewer preferences and past trends, ultimately enhancing discoverability, tenant engagement, and rental efficiency.

Keywords: Sentiment Analysis, Natural Language Processing (NLP), Rental Market, Predictive Modeling, Tenant Engagement.

1. INTRODUCTION

The integration of Natural Language Processing (NLP) and Machine Learning (ML) in India's rental housing sector addresses critical gaps in discoverability, personalization, and data-driven decision-making.

Fig.1 AI in real estate

By enhancing property visibility through optimized content and tailoring listings to match tenant sentiment, landlords can significantly boost engagement and reduce vacancy periods. Machine learning enables accurate, dynamic pricing strategies, while automated sentiment analysis reduces reliance on brokers, streamlining the rental process and lowering costs. These technologies also allow landlords to refine marketing strategies, improve tenant satisfaction, and accelerate rental transactions through actionable insights gathered from user feedback, reviews, and social media conversations. Applications span rental platforms like NoBroker and 99acres, smart property management dashboards, AI-driven chatbots, sentiment monitoring on social media, and automated feedback systems that help owners adapt quickly to market needs. Additionally, dynamic pricing engines and tenant-matching algorithms improve competitiveness by analyzing emotional and practical preferences, ensuring listings meet specific renter demands. Sentiment-based real estate investment analysis further assists investors in identifying high-interest areas and optimizing portfolios. Urban planners and policymakers can also benefit by using aggregated sentiment data to design better housing policies and infrastructure projects. Educational dashboards and landlord training tools provide a foundation for understanding

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

and applying these insights effectively. Overall, the proposed system promotes transparency, personalization, efficiency, and smarter decision-making, transforming the rental ecosystem into a more intelligent, user-centric, and technologically advanced environment.

2. LITERATURE SURVEY

Armstrong et al.'s [2] advice, a review of prior knowledge must be carried out before constructing a formidable forecasting model. The years of causal inference can contribute important insights and help avoid nonsensical relationships that models sometimes assign by chance, thus, to obtain a solid theoretical basis for the forecasting model, a literature review analysis was conducted in three parts. First, a review of previously used variables and their effects on price was carried out, which directed choosing candidate variables in the forecasting model. The second step examined scientific studies that attempted to measure variable importance, emphasising the literature gap. Finally, in the third step, the review of real estate and pandemic studies was discussed to gather any additional insight that could be helpful for model explanation or construction.

The first variable on the list was the most intriguing and widely discussed covariate among real estate scholars: the so-called TOM variable. The best summary of this variable's effect can be described via the study completed by Benefield et al. [9], where out of 197 price equation estimations, 73 instances reported insignificant, 24—positive and 100—negative TOM relationships with the real estate price. These findings stem from two long-established theories: the search theory formed by Yinger [10] and the sale clearance theory of Lazear [11].

The former theory states that the longer a property is on the market (listed on the real estate website), the higher the probability is to discover a buyer that is willing to pay the highest price. This notion intuitively makes sense, as not all buyers are constantly refreshing websites and spotting every single property in the sea of listings. As full-time work and other personal matters consume most time for any individual, a longer TOM does not necessarily increase the likelihood of a price drop but inversely helps to find a buyer willing to pay the highest price.

In contrast, the Lazear [12] clearance model states that high TOM values for a property simply indicate a lack of buyer interest, thus, to make the property more attractive, the price needs to be reduced. The authors who sympathise with this theory argue that with longer TOM values, a certain stigma is attached to the property, as if it is not valuable or something is inherently wrong with it. The most recent papers by An et al. [3] and He et al. [13] further attempted to explain the TOM phenomenon. An et al. [14] claimed that the TOM effect on the price solely depends on the market conditions, meaning that in times of high growth, a longer TOM should help find the best buyer, but in times of economic downfall, higher TOM values will negatively affect the selling price. He et al. [15] argued that the TOM relationship is non-linear and possesses an inverted U-shaped component, meaning that up to a certain point, the TOM variable raises the chance of finding the best buyer, but after the inflection point, the TOM effect becomes negative.

Two points regarding the TOM variable must be considered. First, most of the studies tried to establish a linear model, which confines the dynamics of the TOM variable. Second, researchers have used different local market datasets. It could be that geographical locations exhibit different results. Either way, due to many differing conclusions, it is cumbersome to grasp the magnitude or the direction of the TOM variable effect while relying on earlier studies. Nonetheless, many papers consider the TOM variable an important factor influencing real estate prices; therefore, this variable is essential in the forecasting model.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

The empirical findings provided by Huang and Palmquist [4], Knight [5], Anglin et al. [1], Herrin [6], Johnson et al. [7], Benefield et al. [6] and Verbrugge et al. [8] suggested that the initial price setup or the degree of overpricing can affect the price change. The idea here is that asset owners set an initial price too high with respect to other similar properties on the market and eventually have to reduce their price. This relates to information asymmetry and is acknowledged by many authors; thus, the price variable should also be included.

3. PROPOSED SYSTEM

The proposed system leverages machine learning and Natural Language Processing (NLP) techniques to analyze Airbnb listings and predict viewer interest and listing performance. The process begins with uploading a comprehensive dataset containing listing attributes, host information, pricing, availability, and customer reviews.

This data is then preprocessed by handling missing values, encoding categorical variables, and applying NLP methods such as stopword removal, tokenization, and TF-IDF vectorization on text-heavy columns like "description" and "neighborhood_overview." These preprocessing steps ensure the dataset is clean, consistent, and ready for model training.

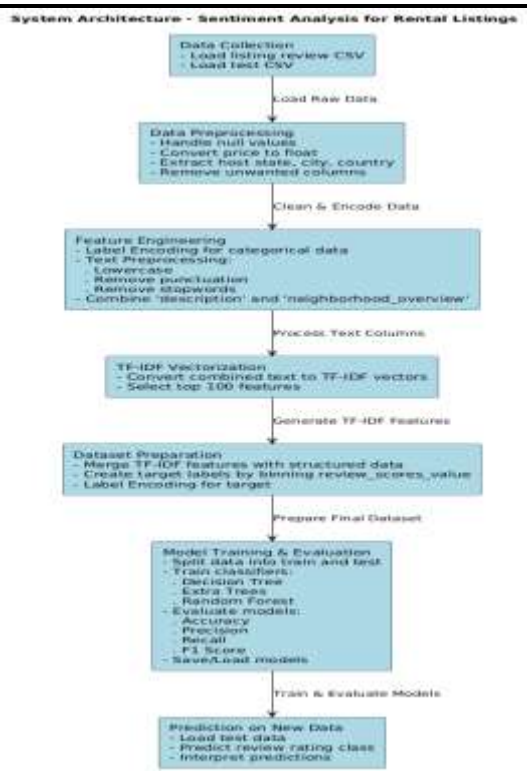

Fig.2: Block Diagram

Initially, the Extra Trees Classifier is used as a baseline model. This ensemble learning technique builds multiple randomized decision trees and combines their outputs for robust classification.

Though efficient and resistant to noise, the Extra Trees Classifier may not always offer the best accuracy due to its high randomness in feature selection and lack of bootstrap sampling.

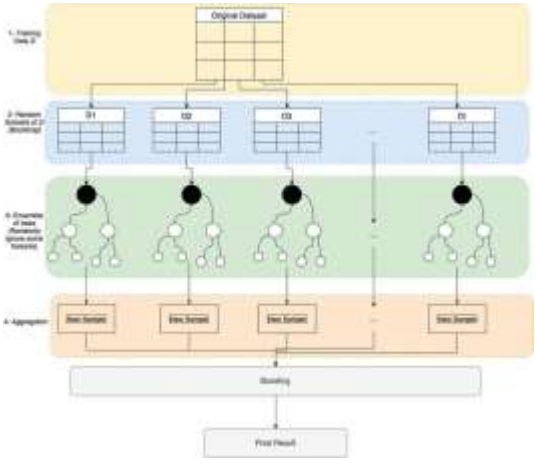

Fig.3: ETC Model Architectural Diagram

To address this, the system proposes the use of the Random Forest Classifier, which improves

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

performance by employing bootstrap sampling and random feature selection at each split. This approach results in more accurate, stable, and generalizable predictions across the dataset.

Data splitting is conducted using an 80-20 ratio to create training and testing sets. The feature set includes structured fields and numerical vectors from text-based columns. Model training begins with fitting the Extra Trees Classifier and evaluating its performance using metrics such as accuracy, precision, recall, and F1-score. Subsequently, the Random Forest Classifier is trained and optimized using GridSearchCV for hyperparameter tuning. This classifier typically outperforms the baseline model due to its better handling of variance and more balanced bias-variance tradeoff.

The architecture of the Extra Trees Classifier involves training multiple decision trees on the full dataset, introducing randomness at each split without using bootstrap samples. While computationally efficient, it may lead to information loss due to random feature selection.

In contrast, the Random Forest Classifier uses bootstrap samples and random feature selection to build a diverse set of decision trees, aggregating their predictions through majority voting. This reduces overfitting and improves generalization on unseen data.

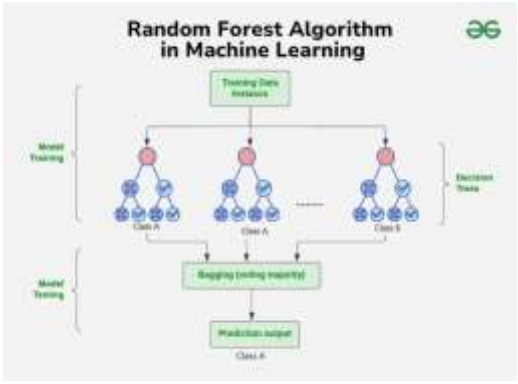

Fig.4: RFC Model Architectural Diagram

Advantages of the Random Forest Classifier include higher accuracy, effective overfitting control, and the ability to handle high-dimensional data and missing values. It also provides feature importance scores, aiding interpretability and feature selection. The trained model is saved using tools like joblib or pickle for future real-time applications, such as predicting pricing or customer sentiment for new listings. This proposed system enhances the automation and intelligence of rental listing platforms, enabling more personalized, data-driven marketing strategies.

4. RESULTS AND DISCUSSION

Figure 5 displays the uploaded dataset used for conducting sentiment analysis on rental property listings and viewer feedback. This dataset comprises structured information where each row represents a unique entry such as a review or rental listing, and each column captures specific attributes related to the entry. Common fields in the dataset include review text, rating scores, property descriptions, location details, pricing, and timestamps. The successful upload of the dataset signifies the initial and essential step toward data preprocessing, exploration, and subsequent application of natural language processing and machine learning techniques. It provides a foundation for extracting meaningful insights related to viewer sentiments and their correlation with rental property features.

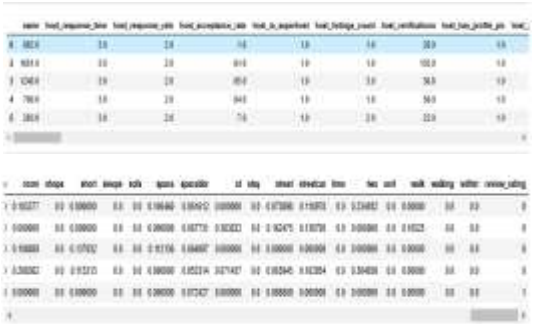

Fig.5: Uploaded Dataset

Figure 6 presents the count distribution of review score ratings contained within the dataset. It visually illustrates how frequently each rating level appears, thereby offering insights into overall viewer sentiment. Ratings are likely categorized on a numerical scale—such as from 1 to 5—and the frequency of each score is plotted to assess the sentiment polarity within the reviews. A higher

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

frequency of ratings like 4 or 5 indicates predominantly positive sentiment, while a significant number of lower scores such as 1 or 2 reflects dissatisfaction or negative sentiment. This figure helps in understanding the sentiment distribution and is essential for building and validating sentiment classification models, ensuring balanced and representative input for training.

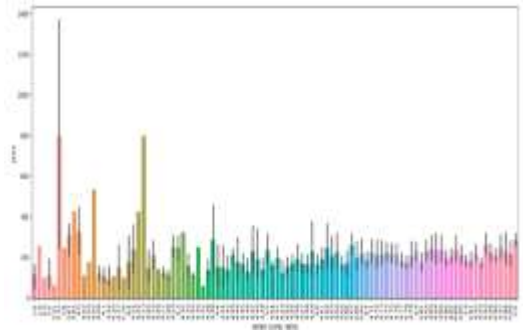

Fig.6: Count of Review Score rating

```
Model saved successfully.
ExtraTreesClassifier Accuracy    : 74.67811158798283
ExtraTreesClassifier Precision   : 50.0
ExtraTreesClassifier Recall      : 37.33905579399141
ExtraTreesClassifier FSCORE      : 42.75184275184275

 ExtraTreesClassifier classification report
              precision    recall  f1-score   support

     Ordinary       0.75      1.00      0.86       174
Extraordinary       0.00      0.00      0.00        59

     accuracy                           0.75       233
    macro avg       0.37      0.50      0.43       233
 weighted avg       0.56      0.75      0.64       233
```

Fig.7: Classification of ETC

The Extra Trees Classifier achieved an accuracy of **74.68%**, with a precision of **50.0%**. The recall was measured at **37.34%**, indicating the model's ability to identify relevant instances, while the F1-score stood at **42.75%**, reflecting the balance between precision and recall.

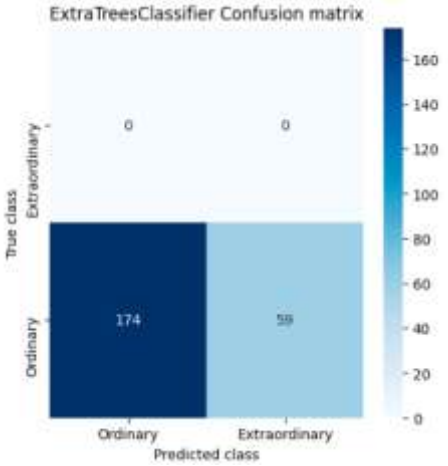

Fig.8: ETC Confusion Matrix

```
Model saved successfully.
RandomForestClassifier Accuracy    : 85.83690987124464
RandomForestClassifier Precision   : 77.07481005260081
RandomForestClassifier Recall      : 83.42352002352093
RandomForestClassifier FSCORE      : 79.43512797881729

 RandomForestClassifier classification report
              precision    recall  f1-score   support

     Ordinary       0.87      0.95      0.91       174
Extraordinary       0.80      0.59      0.68        59

     accuracy                           0.86       233
    macro avg       0.83      0.77      0.79       233
 weighted avg       0.85      0.86      0.85       233
```

Fig.9: Performance Evaluation of RFC

Figure 9 shows that The Random Forest Classifier demonstrated strong performance with an accuracy of 85.84%, indicating a high overall correctness in predictions. It achieved a precision of 77.07%, meaning that 77.07% of the positive predictions were correct. The recall stood at 83.42%, showing the model's ability to identify relevant instances, while the F1-score of 79.44% highlights a good balance between precision and recall, making it a reliable classification model.

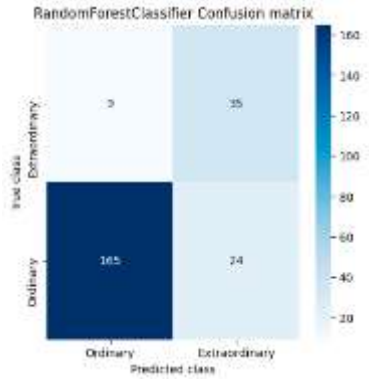

Fig.10: Confusion Matrix

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

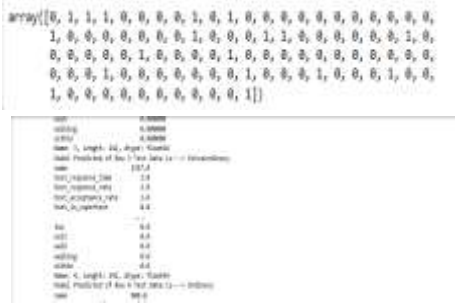

Fig.11: Model Prediction on Test Data.

The model's predictions for the test data indicate that **Row 3** is classified as "**Extraordinary**," while **Row 4** is classified as "**Ordinary**."

5. CONCLUSION

The integration of Natural Language Processing (NLP) and machine learning into rental property marketing offers a data-driven approach to understanding viewer sentiment and optimizing property listings. Traditional rental marketing methods have relied heavily on intuition, brokers, and manual feedback, often resulting in inefficiencies and increased costs. By applying sentiment analysis, landlords can gain deeper insights into tenant preferences, leading to improved property descriptions, more accurate pricing strategies, and more effective marketing efforts. Machine learning models classify customer sentiment based on listing descriptions, inquiries, and reviews, enabling landlords to predict viewer interest with greater precision. This approach enhances property visibility, increases tenant engagement, and streamlines rental transactions through personalized recommendations. Moreover, automated feedback analysis reduces dependence on brokers, resulting in more targeted and efficient listings. This study demonstrates how predictive models can revolutionize the rental market by providing actionable insights that ensure properties are marketed to the right audience. In conclusion, NLP-driven sentiment analysis offers a significant competitive edge in the evolving rental landscape, empowering landlords to make informed, data-backed decisions that improve occupancy rates and optimize rental pricing.

REFERENCES

[1] Anglin PM, Rutherford R, Springer TM. The trade-off between the selling price of residential properties and time-on-the-market: the impact of price setting. J Real Estate Finan Econ. 2003;26(1):95–111.

[2] Armstrong JS, Green KC, Graefe A. Golden rule of forecasting: be conservative. J Real Estate Res. 2015;68(8):1717–31. https://doi.org/10.1016/j.jbusres.2015.03.031.

[3] AnZ, Cheng P, Lin Z, Liu Y. How do market conditions impact the price-TOM relationship? Evidence from real estate owned (REO) sales. J Hous Econ. 2013;22(3):250–63. https://doi.org/10.1016/j.jhe.2013.07.003.

[4] Ambrus A, Field E, Gonzalez R. Loss in the time of cholera: long-run impact of a disease epidemic on the urban landscape. Am Econ Rev. 2020;110(2):475–525.

[5] Baldominos A, Blanco I, Moreno JA, Iturrarte R, Bernárdez O, Afonso C. Identifying real estate opportunities using machine learning. Appl Sci. 2018;8(11):1–23.

[6] Benefield JD, Cain CL, Johnson KH. On the relationship between property price, time-on-market, and photo depictions in a multiple listing service. J Real Estate Finan Econ. 2009;43(3):401–22.

[7] Berawi MA, Miraj P, Saroji G, et al. Impact of rail transit station proximity to commercial property prices: utilizing big data in urban real estate. J Big Data. 2020; 7:71. https://doi.org/10.1186/s40537-020-00348-z.

[8] Brownlee J. Machine learning mastery with Python. 2020; Ebook.

[9] Benefield J, Cain C, Johnson K. A review of literature utilizing simultaneous modelling techniques for property price and time-on-market. J Real Estate Lit. 2014;22(2):149–75.

[10] Borde S, Rane A, Shende G, Shetty S. Real estate investment advising using machine learning. Int Res J Eng Tech (IRJET). 2017;4(3):1821–5.

[11] Bogin A, Doerner W, Larson W. Local house price dynamics: new indices and stylized facts. Real Estate Econ.
2018. https://doi.org/10.1111/1540-6229.12233.

[12] Buckland M, Gey F. The relationship between recall and precision. JASIST. 1994;45(1):12–9. https://doi.org/10.1002/(sici)1097-4571(199401)45:1%3c12::aid-asi2%3e3.0.co;2-l.

[13] Christoph M. Interpretable machine learning. A Guide for Making Black Box Models Explainable. 2019. https://christophm.github.io/interpretable-ml-book/. Accessed 20 Dec 2020.

[14] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell. 2011;2011(16):321–57.

[15] Chawla N.V. (2009) Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook. Springer, USA. https://doi.org/10.1007/978-0-387-09823-4_45.