International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 15, Issue 5, May 2026)

An Efficient Machine Learning Technique for Fake Review Prediction on Amazon Dataset

[1]Nidhi Sharda, [2]Pradeep Pal
[1]M. Tech Scholar, [2]Assistant Professor
Department of Computer Science and Engineering
Lakshmi Narain College of Technology, Indore, India

***Abstract*— Fake reviews have become a significant challenge for e-commerce platforms, as they can mislead customers and negatively affect product credibility and purchasing decisions. This study presents an efficient machine learning technique for fake review prediction on the Amazon dataset using Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) classifiers. The review data are preprocessed through cleaning, tokenization, and feature extraction to improve the quality of input information. The developed framework analyzes review content and relevant attributes to distinguish genuine reviews from deceptive ones. The performance of the selected machine learning models is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Comparative analysis demonstrates the strengths and limitations of each classifier in identifying fraudulent reviews. The results indicate that machine learning-based classification can effectively detect fake reviews and enhance the reliability of online review systems. The proposed approach contributes to improving consumer trust, supporting informed purchasing decisions, and maintaining the integrity of Amazon's e-commerce ecosystem.**

Keywords—Machine learning, E- Commerce, Python, Accuracy, Error rate.

I. INTRODUCTION

The rapid advancement of internet technologies and the widespread adoption of e-commerce platforms have significantly transformed consumer purchasing behavior. Online shopping has become an integral part of daily life, enabling customers to purchase products and services conveniently from anywhere in the world. Among the numerous e-commerce platforms available today, Amazon is one of the most popular and influential online marketplaces, offering millions of products across diverse categories. As the number of online transactions continues to grow, customers increasingly rely on digital information sources to make informed purchasing decisions. One of the most important sources of such information is the customer review system, which allows buyers to share their experiences, opinions, and satisfaction levels regarding purchased products.

Customer reviews play a vital role in influencing consumer decisions because they provide valuable insights into product quality, functionality, durability, and overall user experience. Before purchasing a product, potential customers often examine ratings and reviews to evaluate whether the product meets their expectations. Positive reviews generally enhance customer confidence and contribute to higher sales, while negative reviews may discourage potential buyers from making purchases. Consequently, online reviews have become a powerful factor affecting product popularity, brand reputation, and seller credibility within the digital marketplace.

Despite their importance, online review systems are increasingly vulnerable to manipulation through fake reviews. Fake reviews are intentionally created misleading opinions that do not accurately reflect genuine customer experiences. These reviews may be posted by sellers attempting to improve product ratings, competitors seeking to damage rival products, or third-party agencies hired to influence customer perceptions. Such deceptive reviews distort the authenticity of online feedback systems and create difficulties for consumers attempting to identify trustworthy products. As a result, fake reviews have become a major concern for e-commerce companies, consumers, and researchers alike.

The problem of fake reviews has grown substantially due to the increasing competition among online sellers.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 15, Issue 5, May 2026)

Product visibility and ranking on e-commerce platforms are often influenced by customer ratings and review counts. Sellers who receive higher ratings generally attract more customers and achieve better sales performance. Consequently, some individuals engage in unethical practices by generating fraudulent reviews to artificially enhance product reputation or negatively impact competitors. These deceptive activities undermine consumer trust and reduce the reliability of online recommendation systems.

Amazon, being one of the largest e-commerce platforms globally, generates enormous amounts of review data every day. Millions of customers continuously provide feedback regarding products ranging from electronics and household items to books and personal care products. The vast quantity of user-generated content makes manual verification of review authenticity nearly impossible. As a result, automated techniques are required to identify suspicious reviews and maintain the integrity of the review ecosystem. The Amazon review dataset has therefore become a valuable resource for researchers investigating review authenticity, consumer behavior, sentiment analysis, and fraud detection.

Figure 1: Artificial Intelligence & E-commerce

Fake review prediction refers to the process of identifying and classifying deceptive reviews within large collections of online feedback. The primary objective is to distinguish genuine customer opinions from manipulated or fraudulent reviews. Accurate prediction of fake reviews can help consumers make more reliable purchasing decisions while protecting businesses from unfair competition. Furthermore, identifying deceptive reviews contributes to maintaining transparency and trustworthiness within e-commerce environments. As online marketplaces continue to expand, the demand for effective fake review prediction systems becomes increasingly important.

One of the major challenges associated with fake review prediction is the evolving nature of fraudulent activities. Individuals who generate fake reviews continuously modify their writing styles, review patterns, and behavioral characteristics to avoid detection. Some deceptive reviews are highly sophisticated and resemble genuine customer feedback, making identification difficult. Additionally, fake reviewers may use multiple accounts, automated tools, or coordinated review campaigns to manipulate product ratings. These factors increase the complexity of detecting fraudulent content and require advanced analytical approaches for effective prediction.

Another challenge arises from the large volume and diversity of review data available on Amazon. Reviews vary significantly in length, language style, sentiment, and content quality. Some reviews contain detailed product descriptions and personal experiences, while others consist of only a few words. Moreover, customer behavior differs across product categories, geographical regions, and demographic groups. This diversity creates additional difficulties in developing generalized prediction systems capable of accurately identifying fake reviews across different scenarios.

The availability of Amazon review datasets has facilitated extensive research in the area of review analysis and prediction. These datasets typically include review text, ratings, timestamps, reviewer information, product details, and other attributes that provide valuable insights into customer behavior. Researchers utilize these datasets to study patterns associated with genuine and deceptive reviews, identify behavioral indicators of fraud, and evaluate the effectiveness of prediction models. The rich information contained within Amazon datasets makes them one of the most widely used resources for fake review research.

Machine learning has emerged as an effective solution for addressing the challenges associated with fake review prediction. By learning patterns from historical review data,

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 15, Issue 5, May 2026)

machine learning models can automatically classify reviews and identify suspicious content. Popular machine learning approaches such as Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) have demonstrated promising results in various classification applications. These techniques analyze review-related features and generate predictive models capable of distinguishing genuine reviews from deceptive ones with high accuracy. Their ability to process large-scale datasets and adapt to complex data patterns makes them particularly suitable for fake review detection tasks.

The application of machine learning to fake review prediction offers several advantages over traditional manual verification methods. Automated systems can process thousands of reviews within a short period, reduce human effort, and provide consistent classification performance. Additionally, machine learning techniques can uncover hidden relationships and subtle patterns that may not be apparent through manual inspection. These capabilities contribute to more efficient and reliable detection of fraudulent reviews in large e-commerce platforms.

II. SMETHODOLOGY

The methodology of the proposed research work is as followings-

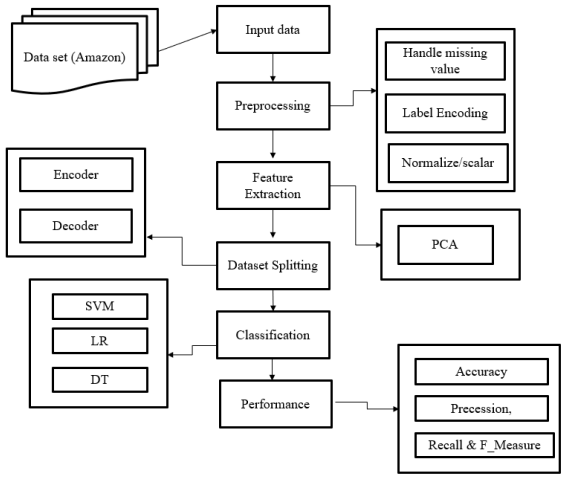

Figure 2: Flow Chart

- **Collect data set**

  The customer review on online product behaviour data set of the Amazon website will be collected from the kaggle machine learning repository for the purpose of putting the research efforts into execution.

  This collection of data contains 69000 customer reviews covering a wide range of items.

- **Processing of data in advance**

  In order to make the evaluation procedure more straightforward, the pre-processing of data comprises transforming any string variables to numerical ones. Take into account missing and null data as well.

  • Dimensionality Reduction Feature extraction is a technique of dimensionality reduction in which an initial collection of raw data is reduced to more manageable groups for the purpose of processing. These groups may then be analysed. While extracting features, keep in mind things like the product name, ratings, and user names.

- **Classification**

  In order to forecast how customers would rate various items, we make use of an algorithm called decision tree.

Algorithm

Input: Customer Behaviour analysis of Reviews of Amazon Products dataset.

Take the initial data features reviews rating, reviews text, reviews title and reviews, username.

Filtering the null value

Classify the text based on sentiments

Output: Optimal Precision, Recall, F-Measure, Accuracy and Error rate

Step: 1. Split train and test dataset Y_train, Y_test, X_train and X_test

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 15, Issue 5, May 2026)

2. Feature extractions, features = {} for word in words: features [word] = True

3. Vectorization

Y train counts

Y train transformer

4. Apply the decision tree machine learning classifier.

5. Generate confusion matrix and show value of TP, FP, TN and FN

6. Calculate Accuracy, error rate, precision, recall and f-measure

7. Plot the ROC Curve

Evaluation

The confusion metrics used to evaluate a classification model are accuracy, precision, and recall.

- Precision = True Positive/(True Positive + False Positive)
- Recall = True Positive/(True Positive + False Negative)
- F1-Score = 2x (Precision x Recall)/(Precision + Recall)
- Accuracy = [TP +TN] / [TP+TN+FP+FN]
- Classification Error = 100- Accuracy

III. SIMULATION & RESULTS

Python is the software that will be used to carry out the simulation. Python is open source programme that has a wide library of artificial intelligence, machine learning, and other related projects. Python, which will be used to construct and simulate the suggested notion, will use the spyder ISE as its platform.

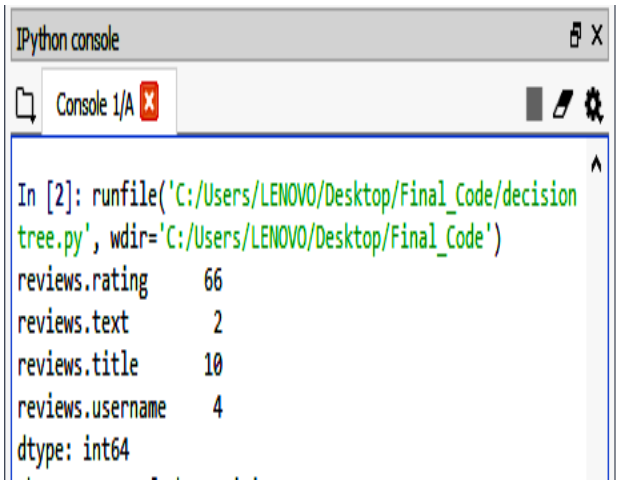

Figure 3: Dataset loading and preprocessing

The online Amazon product review dataset is shown in the python environment in Figure 3, which can be found here. After that, the preprocessing phase begins, during which the following characteristics are extracted: reviews' ratings, reviews' texts, reviews' titles, users' usernames, and so on.

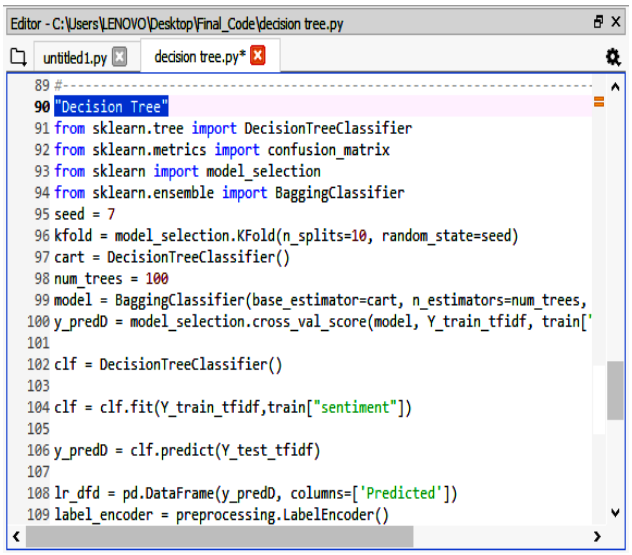

Figure 4: Decision tree classifier

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 15, Issue 5, May 2026)

The decision tree classification method is shown in the python editor window in Figure 4. After the partitioning of the data, the classification technique is carried out. After that, this classifier assigns categories to each of the values in the dataset and either produces a confusion matrix or a projected model.

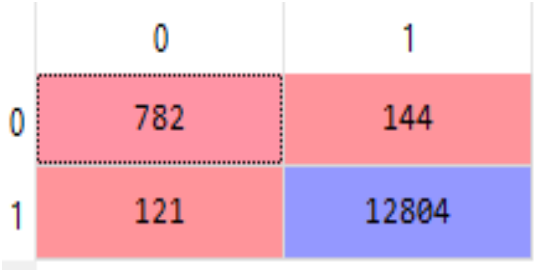

Figure 5: Confusion Matrix (DT)

The predicted value from decision tree method is as followings-

True Positive (TP) = 782

False Positive (FP) = 144

False Negative (FN) = 121

True Negative (TN) = 12804

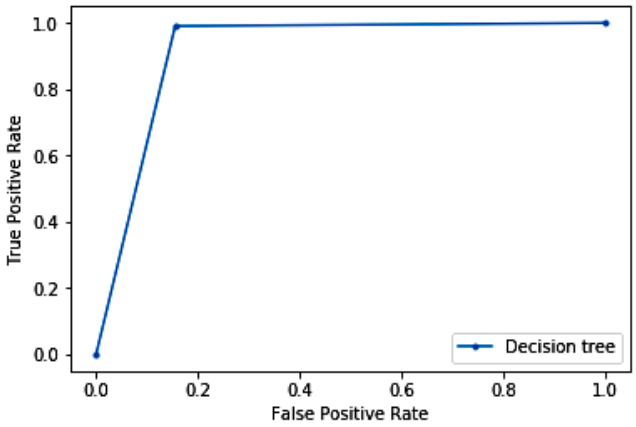

Figure 6: ROC of Decision Tree

The Receiver Operating Characteristic curve is shown in Figure 6, which may be found here (ROC). The True Positive Rate, also known as TPR, may be found on the y-axis, while the False Positive Rate, also known as FPR, can be found on the x-axis.

Table 1: Simulation Result of DT

Sr. No.	Parameters	Value (%)
1	Accuracy	98.11
2	Classification Error	1.89
3	Precision	84.45
4	Recall	86.59
5	F-measure	85.45

Table 1 is showing the simulation results when of the decision tree machine learning classification algorithm.

Table 2: Result Comparison

Sr. No.	Parameters	Previous Work	Proposed Work
1	Method	CNN [1]	Decision Tree
2	Accuracy (%)	97	98.08
3	Classification error (%)	3	1.91
4	Precision (%)	94	95
5	Recall (%)	92	93
6	F-measure (%)	93	95

IV. CONCLUSION

In this work, an efficient machine learning-based fake review prediction framework was developed using Amazon review data. The dataset was preprocessed and analyzed using three machine learning classifiers, namely Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT). The models were trained and evaluated using standard classification measures to determine their

effectiveness in distinguishing genuine and fake reviews. Experimental results demonstrated that the Decision Tree classifier achieved superior performance compared to the previous CNN-based approach. The proposed model attained an accuracy of 98.08%, precision of 95%, recall of 93%, and F-measure of 95%, while reducing the classification error to 1.91%. These outcomes indicate that the proposed approach can effectively identify deceptive reviews and improve the reliability of online review systems. Future research can focus on integrating advanced ensemble and deep learning techniques to further enhance prediction accuracy. Additionally, incorporating reviewer behavioral features and real-time review monitoring mechanisms may improve the detection of increasingly sophisticated fake review patterns.

REFERENCES

1. S. C, R. S and U. K, "Fake Review Detection and Classification Using Improved Convolutional Neural Network on Amazon Dataset," 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2023, pp. 398-403, doi: 10.1109/ICPCSN58827.2023.00071.
2. J. Jeong, "Identifying Customer Preferences From User-Generated Content on Amazon.Com by Leveraging Machine Learning," in IEEE Access, vol. 9, pp. 147357-147396, 2022, doi: 10.1109/ACCESS.2021.3123301.
3. B. Lebichot, T. Verhelst, Y. -A. Le Borgne, L. He-Guelton, F. Oblé and G. Bontempi, "Transfer Learning Strategies for Credit Card Fraud Detection," in IEEE Access, vol. 9, pp. 114754-114766, 2021, doi: 10.1109/ACCESS.2021.3104472.
4. X. Chen, Y. Li, J. Shimada and N. Li, "Online Learning and Distributed Control for Residential Demand Response," in IEEE Transactions on Smart Grid, vol. 12, no. 6, pp. 4843-4853, Nov. 2021, doi: 10.1109/TSG.2021.3090039.
5. S. Wu, W. -C. Yau, T. -S. Ong and S. -C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," in IEEE Access, vol. 9, pp. 62118-62136, 2021, doi: 10.1109/ACCESS.2021.3073776.
6. K. Ali and A. X. Liu, "Monitoring Browsing Behavior of Customers in Retail Stores via RFID Imaging," in IEEE Transactions on Mobile Computing, doi: 10.1109/TMC.2020.3019652.
7. E. Umuhoza, D. Ntirushwamaboko, J. Awuah and B. Birir, "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa," in SAIEE Africa Research Journal, vol. 111, no. 3, pp. 95-101, Sept. 2020, doi: 10.23919/SAIEE.2020.9142602.
8. L. Fan, J. Li and X. -P. Zhang, "Load prediction methods using machine learning for home energy management systems based on human behavior patterns recognition," in CSEE Journal of Power and Energy Systems, vol. 6, no. 3, pp. 563-571, Sept. 2020, doi: 10.17775/CSEEJPES.2018.01130.
9. Y. Yuan, K. Dehghanpour, F. Bu and Z. Wang, "A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand," in IEEE Transactions on Power Systems, vol. 35, no. 5, pp. 4026-4035, Sept. 2020, doi: 10.1109/TPWRS.2020.2979943.
10. F. Zheng and Q. Liu, "Anomalous Telecom Customer Behavior Detection and Clustering Analysis Based on ISP's Operating Data," in IEEE Access, vol. 8, pp. 42734-42748, 2020, doi: 10.1109/ACCESS.2020.2976898.
11. Y. Zhang, S. He, S. Li and J. Chen, "Intra-Operator Customer Churn in Telecommunications: A Systematic Perspective," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 948-957, Jan. 2020, doi: 10.1109/TVT.2019.2953605.
12. E. A. E. Dawood, E. Elfakhrany and F. A. Maghraby, "Improve Profiling Bank Customer's Behavior Using Machine Learning," in IEEE Access, vol. 7, pp. 109320-109327, 2019, doi: 10.1109/ACCESS.2019.2934644.