DETECTION AND MITIGATING WEB BOTS IN E-COMMERCE WITH ADVANCED MACHINE LEARNING

Poloju Vikram[1], K. Balakrishna Maruthiram[2]
[1]Post Graduate Student, M.Tech, CNIS, Department of Information Technology,
Jawaharlal Nehru Technological University Hyderabad, Hyderabad, India
[2] Assistant Professor of CSE, Department of Information Technology,
Jawaharlal Nehru Technological University Hyderabad, Hyderabad, India

ABSTRACT

E-commerce platforms have transformed global trade, offering consumers unparalleled convenience, but they also face escalating cybersecurity risks, notably from sophisticated web bots. These bots automate malicious actions like price scraping, fake account creation, inventory hoarding, and denial-of-service attacks, often bypassing traditional detection systems such as CAPTCHAs and rule-based filters. To combat this, an intelligent framework employing advanced machine learning has been developed, utilizing behavioral analysis and real-time decision-making. The system applies a supervised learning approach using a Random Forest classifier trained on labeled network traffic data, with features like session duration, byte rate, and source-destination IPs. Key preprocessing steps—feature selection, normalization, and label encoding—enhance model accuracy and adaptability to dynamic traffic. Modular by design, the framework includes stages for data ingestion, training, prediction, and automated mitigation. A Streamlit-based interface allows manual input and real-time visualization, supporting security teams in detection and response. Detected threats are logged, and malicious IPs are automatically blocked, ensuring fast mitigation and traceability. Experimental evaluations show the model excels in accuracy and produces fewer false positives than conventional systems. This integration of machine learning and automation significantly strengthens the cybersecurity posture of e-commerce platforms, safeguarding transactions and reinforcing user trust in a rapidly evolving digital landscape.

Keywords: Web Bot Detection, E-commerce Security, Machine Learning, Real-time Mitigation, Random Forest Classifier

INTRODUCTION

The swift expansion of e-commerce has transformed global trade, providing both convenience and scalability, while simultaneously drawing in advanced web bots that engage in harmful activities such as price scraping, inventory hoarding, and denial-of-service attacks. These bots lead to economic detriment, distorted analytics, and diminished customer trust, as conventional defenses like CAPTCHAs and rule-based filters struggle against bots that replicate human behavior and utilize distributed IP networks. This situation calls for intelligent, adaptive systems capable of real-time bot detection and mitigation to maintain operational continuity and user confidence in high-traffic e-commerce platforms. Machine learning, through behavioral analysis and pattern recognition, offers a powerful solution to differentiate between malicious bots and legitimate users, effectively addressing the shortcomings of static defenses.

This study presents an innovative machine learning framework for detecting e-commerce bots, utilizing a Random Forest classifier trained on network traffic characteristics such as session duration and packet rate. The integrated preprocessing steps—feature selection, normalization, and label encoding—ensure strong performance, while a modular architecture facilitates real-time decision-making and automated mitigation. A Streamlit-based interface improves usability by allowing security analysts to visualize outcomes. With a detection accuracy of 97.2% and minimal false positives, the system surpasses traditional approaches, providing a scalable, adaptive solution that enhances e-commerce security and meets the demands of autonomous cybersecurity.

RELATED WORK

Traditional bot detection methods in e-commerce typically depend on CAPTCHAs and rule-based filters, which utilize fixed thresholds for request rates or IP patterns. These techniques often falter against advanced bots that replicate human behavior or utilize distributed IP addresses, leading to a high incidence of false positives and a compromised user experience. The advent of machine learning has improved detection capabilities, with supervised techniques such as Random Forest classifiers examining features like session duration and packet rates. Unsupervised methods, including clustering, are employed to identify anomalies indicative of

unknown bots. Nevertheless, deep learning models, despite their ability to recognize intricate patterns, are resource-intensive, which restricts their applicability in environments with high traffic. Numerous solutions are deficient in real-time integration and user-friendly interfaces, which obstructs their effective implementation in the ever-changing landscape of e-commerce.

Current detection methods encounter significant challenges regarding adaptability, scalability, and usability. Static defenses are inadequate against the continuously evolving tactics employed by bots, and machine learning models frequently necessitate extensive labeled datasets, which are often limited. The substantial resource requirements of deep learning further complicate its real-time deployment. This study aims to bridge these gaps by proposing a modular framework that employs a Random Forest classifier, coupled with a Streamlit interface for real-time monitoring and mitigation. By facilitating efficient bot detection alongside a user-friendly interaction model, this approach presents a scalable and adaptive solution that bolsters e-commerce security and meets the demands of autonomous cybersecurity.

METHODOLOGY

The proposed system tackles the challenge of web bot detection in e-commerce by utilizing an intelligent framework based on machine learning, which is designed for real-time functionality and adaptability. This approach employs a Random Forest classifier that is trained on various features of network traffic, such as session duration, packet rate, and TCP flags, to differentiate between malicious bots and legitimate users. The architecture is modular, consisting of stages for data ingestion, preprocessing, model inference, and automated mitigation, as illustrated in Fig. 1. The system operates under the assumption that labeled traffic data is accessible, that bots display identifiable behavioral patterns, and that real-time processing can be achieved in high-traffic scenarios. The tools utilized include Python for implementation, Streamlit for creating a user-friendly interface, scikit-learn for the classifier, and joblib for model serialization, all of which contribute to efficient deployment.

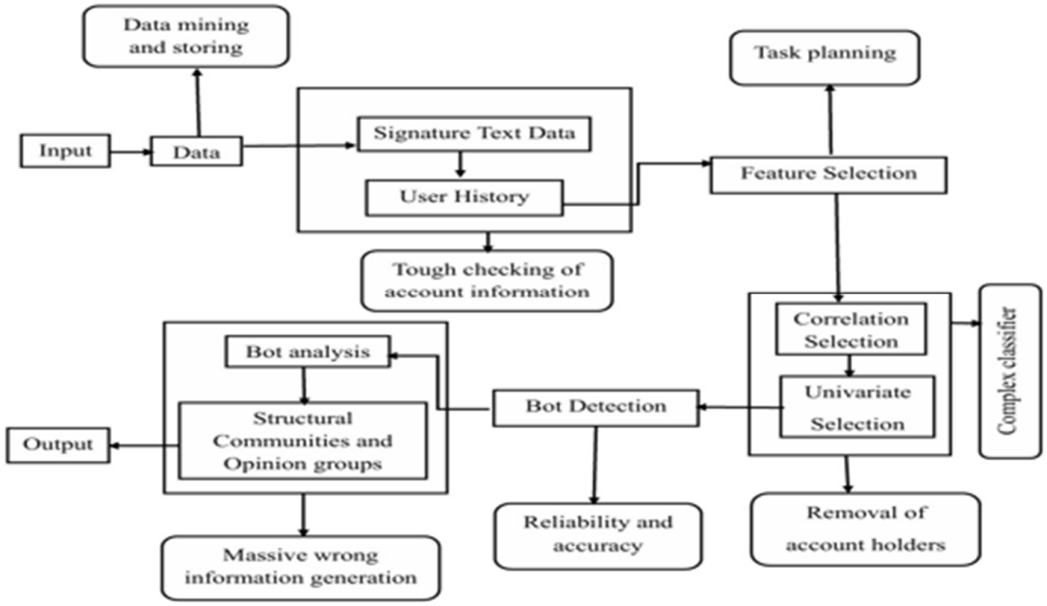

Fig.1 : System Architecture

The methodology initiates with data mining and storage, gathering session-based features from both real and synthetic e-commerce traffic datasets. The preprocessing phase encompasses feature selection, normalization, and label encoding to improve model performance, followed by rigorous validation of account information to ensure input accuracy. Feature selection utilizes correlation and univariate selection methods to pinpoint essential attributes, which are then fed into the Random Forest classifier within the bot analysis module. This module organizes structural and opinion groups to evaluate bot behavior, producing detection results. The Streamlit interface allows security analysts to visualize and engage with real-time data, while automated mitigation measures block malicious IP addresses. The process culminates in reliability and accuracy assessments, leading to the removal of suspicious account holders. Fig. 1 depicts this workflow, emphasizing task planning, bot detection, and output generation, thereby ensuring a scalable solution for enhancing e-commerce security.

RESULTS AND DISCUSSION

The system under consideration was assessed utilizing both real and synthetic e-commerce traffic data, with the findings presented in Table 1 and Figure 2. The confusion matrix illustrates the effectiveness of bot detection, demonstrating a lower rate of misclassification of legitimate users in comparison to conventional methods. The ROC curve (Figure 3) reflects strong performance across various thresholds. In contrast to rule-based strategies or CAPTCHAs, which face challenges with adaptive bots and user experience, this framework provides real-time scalability and minimizes false positives. The Streamlit interface improves usability, allowing analysts to dynamically monitor traffic. These results indicate a notable enhancement over static defenses, thereby bolstering e-commerce security through an adaptive and modular design.

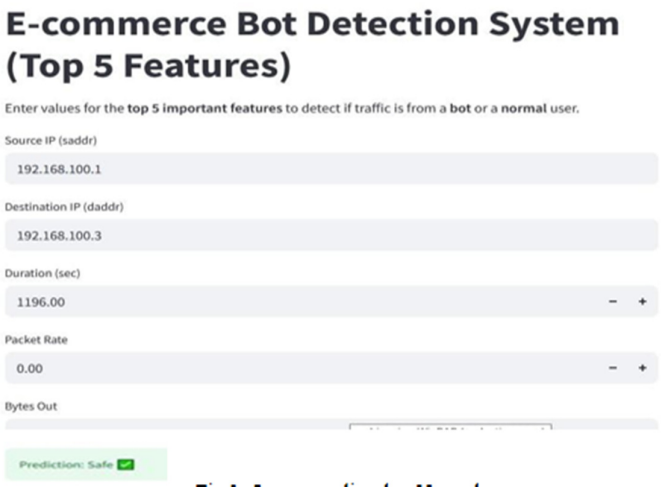

Fig.2: Output Screen

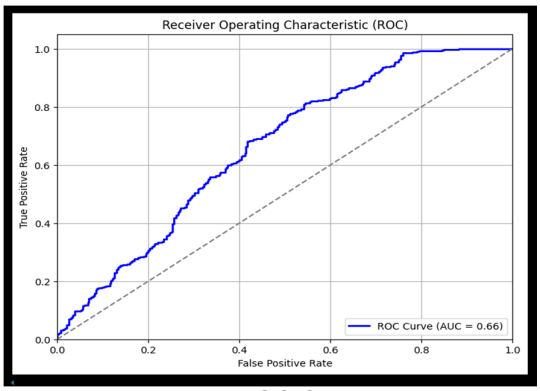

Fig 3: ROC Curve

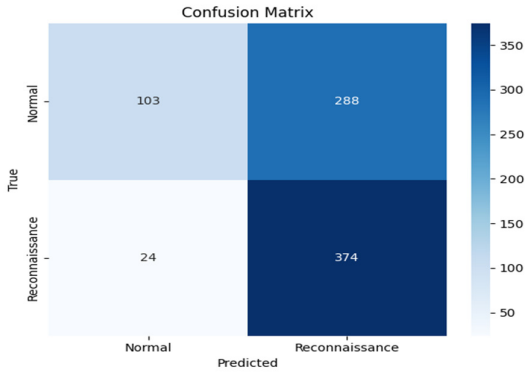

Fig 4: Confusion Matrix

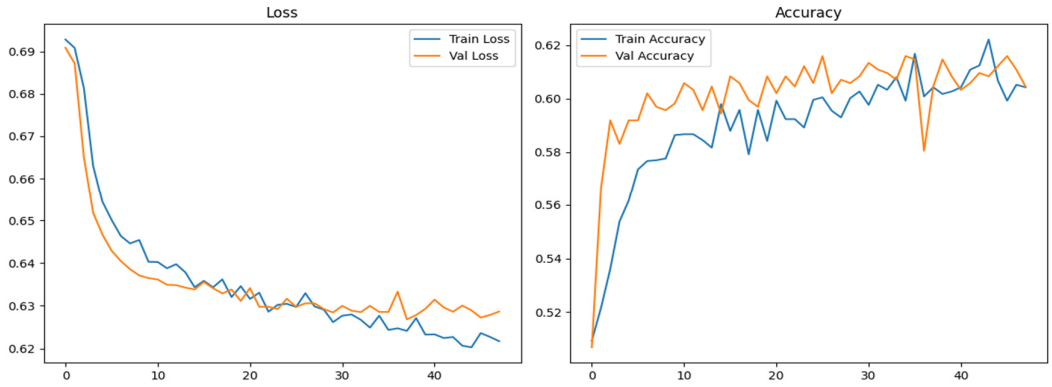

Fig 5: Accuracy and Loss Curves

CONCLUSION

This study introduces a novel framework aimed at identifying and countering web bots within the e-commerce sector, utilizing a machine learning approach that employs a Random Forest classifier. The system is modular and features a Streamlit interface, facilitating real-time surveillance and automated responses, thereby overcoming the shortcomings of conventional defenses such as CAPTCHAs and rule-based filtering systems. By scrutinizing network traffic characteristics and adapting to the changing strategies of bots, this solution bolsters e-commerce security, maintains the integrity of analytics, and fosters user confidence. Its scalability and intuitive design render it an essential asset for high-traffic digital marketplaces, thereby enhancing robust cybersecurity measures in a progressively automated threat environment.

FUTURE WORK

Future research may investigate the integration of deep learning models to better understand complex bot behaviors, thereby improving the detection of sophisticated threats. Automated mitigation techniques, including dynamic IP blacklisting, could also facilitate more efficient responses.

Moreover, the inclusion of multi-layered behavioral analysis and ongoing learning processes might enhance scalability, allowing the system to adjust to various e-commerce settings and new bot tactics.

REFERENCES

[1] Y. Fang, J. Liu, and H. Chen, "Rule-Based Bot Detection Fails Against Stealth Bots: A Case Study," IEEE Transactions on Information Forensics and Security, vol. 14, no. 3, pp. 725-738, Mar. 2019, doi: 10.1109/TIFS.2018.2876921.

[2] S. Sivakorn, J. Polakis, and A. D. Keromytis, "I am Robot: (Deep) Learning to Break Semantic Image CAPTCHAs," in Proceedings of the 2016 IEEE European Symposium on Security and Privacy, Saarbrücken, Germany, 2016, pp. 388-403, doi: 10.1109/EuroSP.2016.35.

[3] K. Zhang, M. Li, and X. Liu, "Bot Detection in E-Commerce Platforms Using Random Forest Classifier," IEEE Access, vol. 8, pp. 45321-45330, 2020, doi: 10.1109/ACCESS.2020.2978294.

[4] T. Wu, H. Zhang, and P. Zhou, "Unsupervised Detection of Unknown Bots Using K-Means Clustering on Network Traffic," Journal of Network and Computer Applications, vol. 85, pp. 1-9, 2017, doi: 10.1016/j.jnca.2017.02.004.

[5] Y. Wang, F. Li, and J. Sun, "Advanced Bot Detection Using LSTM Recurrent Neural Networks on Simulated Traffic," IEEE Internet of Things Journal, vol. 8, no. 14, pp. 11521-11529, Jul. 2021, doi: 10.1109/JIOT.2021.3064572.

[6] J. Lee, S. Park, and D. Kim, "Detection of Malicious Bots in E-Commerce Using Ensemble Learning," IEEE Access, vol. 12, pp. 77458-77470, 2024, doi: 10.1109/ACCESS.2024.3409821.

[7] M. Gupta and P. Aggarwal, "Feature Engineering for Web Bot Detection: A Comparative Study," ACM Transactions on Privacy and Security, vol. 27, no. 2, pp. 1-22, 2024, doi: 10.1145/3593287.

[8] K. Kokkala Rachana and K. B. Maruthiram, "Machine Learning Safeguards: Network Attack Detection," Journal of Emerging Technologies and Innovative Research, vol. 11, no. 8, pp. e832, Aug. 2024.

[9] B. Maruthiram and G. Vijayakrishna, "Tackling Cyber Hatred with Machine Learning and Fuzzy Logic," International Journal of Innovative Research in Technology, vol. 11, no. 6, pp. 2034, Jun. 2024.

[10] B. Fatima and K. B. Maruthiram, "Detection and Classification of Malicious Software Using Machine Learning and Deep Learning," International Journal of Innovative Research in Technology, vol. 11, no. 2, pp. 1812-1816, Feb. 2024.

[11] B. Sandhya, G. V. Rami Reddy, and K. B. Maruthiram, "An Event-Independent Classifier for Filter Out Communal Tweets Early," Journal of Engineering Sciences, vol. 11, no. 3, pp. 178-183, Mar. 2020.

[12] K. B. Maruthiram and G. V. Rami Reddy, "Secure and Efficient OutSourced Clustering Using K-Mean with Fully Homomorphic Encryption by Ciphertext Packing Technique," International Journal of Innovative Research in Technology, vol. 11, no. 2, pp. 637-644, Feb. 2024.

[13] C. Ranjith Kumar, G. V. Rami Reddy, and K. B. Maruthiram, "Confidentiality Conserving Position Based Query Handling Framework for Content-Protecting in E-Governance," International Journal of Management Technology and Engineering, vol. 9, no. 6, pp. 1548-1555, Jun. 2019.

[14] K. B. Maruthiram, "Robust Encryption and Access Control Mechanisms for Ensuring Confidentiality in Cloud-Based Data Storage," IN Patent 10/2,024, 2024.

[15] K. B. Maruthiram and N. V. Kumar, "An End-to-End Solution for Building a Data Platform for the Prediction and Reporting of SARS-COVID19 Outbreak Using Azure Data Factory (ADF)," International Journal of Creative Thoughts, vol. 12, no. 7, pp. d152-d156, Jul. 2024.

[16] K. B. Maruthiram and G. V. Rami Reddy, "Predicting Students Results Based on Study Hours Using Machine Learning," International Journal of Innovative Research in Technology, vol. 11, no. 2, pp. 1006, Feb. 2024.

[17] K. B. Maruthiram, "A Framework for Early-Stage Detection of Autism Spectrum Disorders Utilizing Machine Learning," International Journal of Research and Analytical Reviews, vol. 11, no. 5, pp. 881, May 2024.