WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Spam Detection in IoT Using Machine Learning

Mohd. Nasair Uddin Khan[1], G Rakesh Reddy[2], Varun Boda[3], Sandeep Reddy[4], M. Sri Krishna[5]

[12,3,4,5] Computer Science & Engineering (B. Tech, JNTUH), Sphoorthy Engineering College (JNTUH)
nasairuddinkhansphoorthy@engg.ac.in[1], grakeshreddy@sphoorthyengg.ac.in[2], bodavarun123@gmail.com[3], kuppireddysandeep562@gmail.com[4], srikrishnamorla@gmail.com[6]

Doi: https://doi.org/10.55248/gengpi.5.0524.1391

ABSTRACT

The Internet of Things (IoT) consists of millions of devices with sensors and actuators connected through wired or wireless channels for data transmission. IoT has seen rapid growth over the past decade, with an anticipated 35 billion devices expected to be connected by 2023. The volume of data generated by these devices will increase significantly in the coming years. This data is not only large in volume but also diverse in nature, with varying quality characterized by time and location dependencies. In such a dynamic environment, machine learning algorithms are essential for ensuring security and access control through biometrics, as well as for detecting anomalies to improve the reliability and safety of IoT systems. However, these algorithms can also be targeted by attackers seeking to exploit vulnerabilities in smart IoT systems. In response to these challenges, this paper introduces a method to secure IoT devices by detecting spam using machine learning. The proposed framework, called Spam Detection in IoT using Machine Learning, evaluates five different machine learning models using various metrics and a comprehensive set of input features. Each model calculates a spam score based on refined input features, which indicates the reliability of IoT devices under different conditions. The REFIT Smart Home dataset is utilized to validate the proposed approach. The results demonstrate that this method is effective compared to existing schemes.

INTRODUCTION

The Internet of Things (IoT) enables real-world objects to connect across geographical boundaries, making privacy and security paramount concerns. IoT applications must tackle security challenges such as intrusions, spoofing, DoS attacks, jamming, eavesdropping, spam, and malware. The security measures required for IoT devices vary depending on the organization's size and type.Security measures are influenced by user behavior and device context, including location, nature, and application. For instance, smart security cameras in a smart organization can gather various parameters for analysis and decision-making. Web-based IoT devices, which are highly prevalent, need particular attention to safeguard data privacy and security.Wearable devices that collect and transmit health data to smartphones must ensure the prevention of information leakage to maintain privacy. It is notable that 25-30% of employees connect personal IoT devices to organizational networks, thereby increasing the risk of attacks. The growing scope of IoT attracts both users and attackers.Machine learning (ML) is emerging as a tool in security scenarios, enabling IoT devices to implement defensive strategies that balance security, privacy, and computational requirements. However, the limited resources of IoT systems pose challenges in assessing network conditions and the status of attacks in real-time.

LITERATURE SURVEY

1. Utilizing Machine Learning for Detecting Unwanted Information in IoT Devices

The primary aim of our unsolicited information detection technique is to protect IoT systems from unauthorized access. Unlike traditional methods that focus on analyzing data across messages, web pages, and emails, our strategy targets IoT devices such as sensors, actuators, smart home appliances, intelligent vehicles, and modern gadgets like Google Glasses. These IoT entities produce large volumes of data in various formats, with quality influenced by temporal and spatial factors, particularly speed. Machine Learning (ML) plays a crucial role in IoT ecosystems by providing robust security, user-friendly interfaces, and reliability, which are essential for the development and operation of smart devices. This paper explores the use of ML techniques to identify unwanted information within IoT devices, thereby enhancing system security and operational efficiency.

2. Enhanced Spam Detection in Smart Home IoT Devices Through Ensemble Learning with Time Series Data

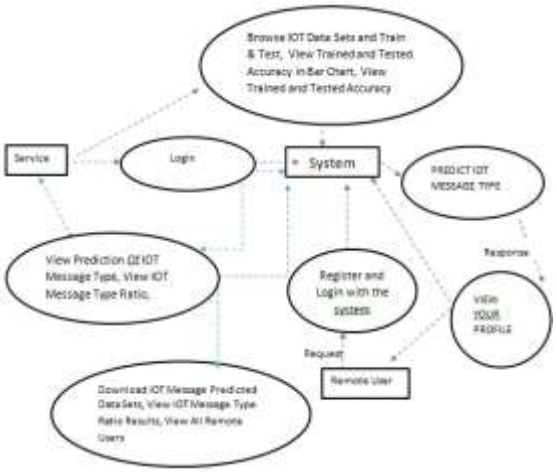
The rise of Internet of Things (IoT) devices in smart homes has led to a significant increase in data generation, primarily transmitted wirelessly. This heightened connectivity also makes IoT devices more susceptible to threats like cyber-attacks and data breaches. This paper aims to improve the security of smart home IoT systems by utilizing statistical analysis and machine learning techniques to detect anomalies in data. We specifically assess the reliability of IoT devices transmitting household appliance readings by analyzing parameters such as feature importance, root mean square error, and hyper-parameter tuning. Using publicly available smart home datasets and weather conditions for validation, our algorithm assigns a spamicity score to each IoT device based on its feature importance and root mean square error score. This score evaluates the trustworthiness of devices within the home network. Our proposed method effectively detects the spamicity score of connected IoT devices, demonstrating its ability to analyze time-series data for spam detection.

3. Advanced Approach for Enhancing Spam Detection in IoT Devices Using Machine Learning

The rapid increase in data generation by Internet of Things (IoT) devices in smart homes, primarily transmitted through wireless channels, presents significant challenges. This data comes in diverse formats and varying quality, influenced by factors such as speed, time, and position dependency. IoT devices are vulnerable to numerous threats, including cyber-attacks, network instability, and data breaches. In this context, machine learning algorithms offer a promising solution for detecting data anomalies and strengthening IoT system security. Our approach focuses on identifying and addressing data anomalies in smart IoT devices, enabling straightforward detection of anomalous events using stored data. We propose an algorithm to determine the spamicity score of connected IoT devices within the network. Experimental results show the effectiveness of our algorithm in analyzing time-series data for efficient spam detection in IoT environments.

PROPOSED APPROACH

In today's digital era, the functionality of smart devices is crucial. Ensuring that the data from these devices is free from spam is imperative. Retrieving information from a variety of IoT devices presents significant challenges due to the diverse domains they operate in. The involvement of numerous devices results in the generation of large volumes of data, collectively known as IoT data, which is characterized by its heterogeneity and variety. This IoT data features real-time updates, multiple sources, and a combination of rich and sparse data.The proposed spam detection system is rigorously validated using five distinct machine learning models.The proposed spam detection system is rigorously validated using five distinct machine learning models.An innovative algorithm is designed to calculate the spamicity score for each model, facilitating effective spam detection and informed decision-making.The spamicity scores are then used to evaluate the reliability of IoT devices through various evaluation metrics, ensuring a thorough and robust analysis.



MODULES

Service Provider Module

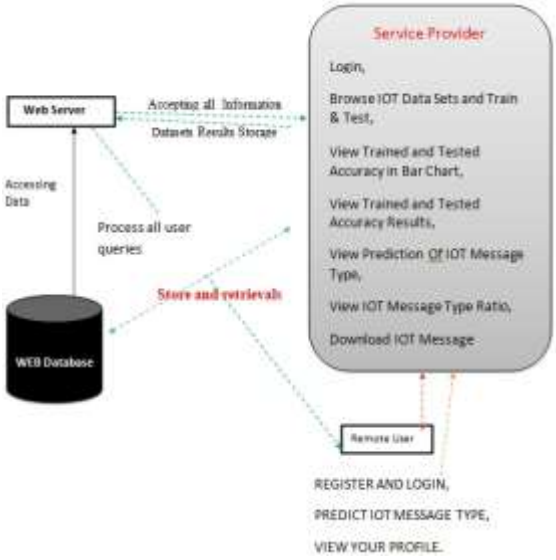
In this module, service providers must log in with valid credentials. After a successful login, they can access a range of functionalities. These include browsing IoT datasets, performing training and testing procedures, viewing accuracy results both graphically and through detailed reports, examining IoT message type predictions, analyzing the distribution of different message types, downloading predicted datasets, and viewing all registered remote users in the system.

View and Authorize Users Module

In this module, administrators can view the list of registered users. They have access to user details such as usernames, email addresses, and physical addresses. Administrators also have the authority to grant or revoke user permissions, thereby managing access to the system's functionalities.

Remote User Module

This module serves the various remote users registered within the system. Users must complete a registration process before engaging in any operations, with their details securely stored in the database. After registering, users need to log in with authorized credentials. Once logged in, users can perform tasks such as registering or logging in, predicting IoT message types, and accessing and editing their profiles.



ALGORITHMS

Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is a highly effective supervised learning algorithm which is designed for classification and regression. It excels at identifying the optimal hyperplane that separates different classes in a multidimensional space. The primary goal of SVM is to maximize the margin between classes, enhancing the model's generalization and robustness to noise. By projecting input data into a higher-dimensional feature space, SVM can effectively manage nonlinear datasets. SVM is also versatile, capable of handling both linearly and non-linearly separable data using kernel functions that transform data into higher dimensions where separation is achievable.

Naive Bayes Algorithm

Despite its simplicity, it is widely used for classification tasks, particularly in text categorization and spam detection. The algorithm assumes that features are independent given the class label, which is why it is termed "naive." However, Naive Bayes often performs well in practice and is computationally efficient, making it suitable for large datasets. It calculates the probabilities of each class based on input features and predicts the class with the highest probability.

Decision Tree Algorithm

The Decision Tree algorithm is a flexible supervised learning tool for classification and regression. It builds a tree-like model by recursively splitting the feature space based on feature values. This process aims to minimize impurity or maximize information gain at each split, resulting in an intuitive and

interpretable tree structure. However, Decision Trees can overfit, particularly with noisy data. Techniques such as pruning and limiting the maximum tree depth help prevent overfitting, ensuring the model remains robust and predictive.

Logistic Regression Algorithm

Logistic Regression is a fundamental algorithm for binary classification, despite its name suggesting a regression method. It predicts the probability of a binary outcome based on independent variables, using the logistic function to map input features to a probability between 0 and 1. Known for its simplicity, interpretability, computational efficiency, and resilience to noise, Logistic Regression is a popular choice for binary classification tasks.

Stochastic Gradient Descent (SGD) Algorithm

Stochastic Gradient Descent (SGD) is a key optimization algorithm used to minimize the loss function in machine learning models. It is particularly well-suited for large-scale and online learning due to its computational efficiency, achieved through incremental parameter updates. SGD iteratively adjusts model parameters in the direction of the steepest descent of the loss function, facilitating rapid convergence and model optimization. Unlike batch gradient descent, SGD uses a stochastic approach, calculating gradients with individual training examples or small batches, which provides exceptional agility in handling large datasets.

CONCLUSION

The proposed framework employs machine learning techniques to identify the spam characteristics of IoT devices. Prior to experimentation, the IoT dataset undergoes preprocessing via feature engineering methods. Through the framework's experimentation with various machine learning models, each IoT appliance is assigned a spam rating. This enhances the criteria necessary for the effective operation of IoT devices within smart home environments. Looking ahead, our future endeavors involve integrating climatic and environmental factors into the analysis of IoT devices, aiming to bolster their security and reliability.

REFERENCES

1) Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, “Iot security: progressing challenges and investigate opportunities,” in 2014 IEEE 7th worldwide conference on service-oriented computing and applications. IEEE, 2014, pp. 230–234.

2) A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, “Blockchain for IoT security and security: The case ponder of a shrewd home,” in 2017 IEEE international conference on unavoidable computing and communications workshops (PerCom workshops). IEEE, 2017, pp. 618–623.

3) E. Bertino and N. Islam, “Botnets and web of things security,”Computer, no. 2, pp. 76–79, 2017.

4) C. Zhang and R. Green, “Communication security in web of thing: preventive degree and maintain a strategic distance from ddos assault over iot network,” in Proceedings of the 18th Symposium on Communications & Organizing. Society for Computer Recreation Universal, 2015, pp. 8–15.

5) W. Kim, O.-R. Jeong, C. Kim, and J. So, “The dim side of the internet: Attacks, costs and responses,” Data frameworks, vol. 36, no. 3, pp. 675–705, 2011.

6) H. Eun, H. Lee, and H. Gracious, “Conditional security protecting security protocol for nfc applications,” IEEE Exchanges on Shopper Electronics, vol. 59, no. 1, pp. 153–160, 2013.

7) R. V. Kulkarni and G. K. Venayagamoorthy, “Neural organize based secure media get to control convention for remote sensor networks,” in 2009 Universal Joint Conference on Neural Systems. IEEE, 2009, pp. 1680–1687.

8) M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, “Machine learning in remote sensor systems: Calculations, techniques, and applications,” IEEE Communications Studies & Instructional exercises, vol. 16, no. 4, pp. 1996–2018, 2014.

9) A. L. Buczak and E. Guven, “A study of information mining and machine learning strategies for cyber security interruption detection,” IEEE Communications Surveys & Instructional exercises, vol. 18, no. 2, pp. 1153–1176, 2015.

10) F. A. Narudin, A. Feizollah, N. B. Anuar, “Evaluation of machine learning for malware detection,” Soft Computing, vol. 20, no. 1, pp. 343–357, 2016.

11) Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, “A system for denial-of-service assault location based on multivariate correlation analysis,” IEEE exchanges on parallel and disseminated frameworks, vol. 25, no. 2, pp. 447–456, 2013.

12) Y. Li, D. E. Quevedo, S. Dey, and L. Shi, “Sinr-based dos assault on remote state estimation: A game-theoretic approach,” IEEE Transactions on Control of Arrange Frameworks, vol. 4, no. 3, pp. 632–642, 2016.

13) L. Xiao, Y. Li, X. Huang, and X. Du, “Cloud-based malware detection game for versatile gadgets with offloading,” IEEE Exchanges on Mobile Computing, vol. 16, no. 10, pp. 2742–2750, 2017.

14) J. W. Department, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, “In-network exception discovery in remote sensor networks,” Knowledge and data frameworks, vol. 34, no. 1, pp. 23–54, 2013.

15) I. Jolliffe, Central component examination. Springer, 2011.

16) A. H. Sodhro, S. Pirbhulal, and V. H. C. de Albuquerque, “Artificial intelligence driven component for edge computing based industrial applications,” IEEE Exchanges on Mechanical Informatics, 2019.

17) R, “Rstudio,” 2019 (gotten to October 23, 2019).

18) L. College, “Refit savvy domestic dataset,” https://repository.lboro.ac.uk/ articles/REFIT Savvy Domestic dataset/2070091, 2019 (gotten to April 26, 2019).

19) A. H. Sodhro, Z. Luo, G. H. Sodhro, M. Muzamal, J. J. Rodrigues, and V. H. C. de Albuquerque, “Artificial insights based qos optimization for interactive media communication in iov systems,” Future Generation Computer Frameworks, vol. 95, pp. 667–680, 2019.

20) I. Guyon and A. Elisseeff, “An presentation to variable and feature selection,” Diary of machine learning investigate, vol. 3, no. Damage, pp. 1157–1182, 2003.

21) L. Yu and H. Liu, “Feature determination for high-dimensional information: A fast correlation-based channel solution,” in Procedures of the 20th international conference on machine learning (ICML-03), 2003, pp. 856–863.