WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Determining Spam Issues in IoT Data Using Machine Learning

K Bhanu Naveen Teja[1], B Sundara Kumar[2], Bala Venkata Seetharam[3], A Phani Sridhar[4]

[4]Professor, [1,2,3] Department of CSE, Aditya Engineering College, Surampalem, AP, India.

ABSTRACT:

Internet of Things (IOT) is composed of a vast number of sensors and actuators interconnected via either wireless or wired channels. The variety of information that this devices produce will increase dramatically in the coming years. IoT devices produce a lot of information generally as well as in a range of various modalities, with varying degrees of data quality depending on the speed of time and positional reliance. The assurance of bioengineering security and authentication, along with anomaly detection to improve the utility and trustworthiness of IoT devices, can be achieved in such a scenario by machine learning algorithms. The idea behind the breakthrough is that a machine can produce precise findings simply by studying from the input. Thus, we suggest that IoT device security be achieved by employing machine learning to detect spam.

Key Word– Random forest, logistic regression, support vector machine, REFIT, smart home, spam.

1. INTRODUCTION

OVERVIEW

IOT enables integration and synchronization among physical objects, regardless of where they are located geographically. Privacy and protection measures become essential and challenging in this case due to the use of various network control and administration technologies. IoT applications must protect user data privacy to deal with security issues such intrusions, mimicking assaults, Denial of service, blocking, snooping, spam, and malware.

It is crucial to make sure that the IOT data is safe to depend on in light of the advancement of ML in order to generate any future discoveries or surveys. Thus, a framework is needed to make sure that the data is free of any spam features that can undermine the reliability of the IOT network.

OBJECTIVE

IoT devices produce a variety of data types while collecting data, such as formatted, semi formatted, and unformatted information. This data may consist of periodic sensor readings, analogue transmissions, details about the health of the device, or substantial image or video files. As IoT data is not homogeneous, there is no single, accepted method for storing it. To be usable, IoT data must be analysed, but doing so manually is challenging given the huge amount of information that IoT devices generate. Automated analytics are thus a critical part of the majority of IoT solutions. On the basis of the telemetry data, analytics tools are used to produce illuminating reports, display data through dashboards and data visualization techniques, and set off alarms and actions. Thus, we put up a framework for IoT device security utilising machine learning to detect spam. The goal is to fix the problems brought on by spam parameters created by attackers that are impairing the reliability of the IOT network.

2. LITERATURE SURVEY

- **IOT security: ongoing challenges and research opportunities**

AUTHORS: Z.-K. Zhang, M.C.Y. Cho, C.W. Wang, C.W. Hsu, C.-K. Chen, S. Shieh

THEME: Internet of Things, software, home appliances, and wearable electronics are now capable of transferring and exchanging data online thanks to the Internet of Things Information security on the shared data is a crucial issue which cannot be avoided because it contains a significant amount of private information. This paper deals with an overview of IoT's general information security history before moving on to the issues that IoT will face in terms of information security. Last but not least, this will point to potential future study fields for the creation of solutions to the security breaches that the IoT faces.

- **Communication security in internet of thing: preventive measure and avoid ddos attack over iot network**

AUTHORS: C. Zhang and R. Green

THEME: IoT security concerns must be appropriately handled because the use of IoT is expanding in many crucial industries. One of the most well-known attacking techniques is Distributed Denial of Service (DDoS), It entails using a cluster of ghost machines linked to the web from multiple locations

to overwhelm the host server with a lot of queries.To examine the interactive connection between various types of network nodes, a simple defensive approach against DDoS attacks in IoT network environments is offered and put into practise in a number of scenarios.

- **Blockchain for IOT security and privacy: The case study of a smart home.**

AUTHORS: A .Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram

THEME: By thoroughly analysing its security in relation to the core security objectives of privacy, integrity, and reliability, this paper concluded that the BC-based smart home framework is secure. The simulation results were then provided to demonstrate that, in comparison to the security and privacy benefits their technique offers, the operating costs (in terms of data transfer, process time, and power consumption) are minimal.

- **Neural network based secure media access control protocol for wireless sensor networks**

AUTHORS: R. V. Kulkarni and G. K. Venayagamoorthy

THEME: This paper addresses the integration of neural networks to wireless sensor network (WSN) security. It provides a media access control protocol (MAC) based on multilayer perceptrons (MLP) to guard against hostile denial-of-service attacks on a CSMA-based WSN. By continuously observing the variables that show odd changes in the event of an attack, the MLP strengthens the security of a WSN. The MLP-protected secure WSN is implemented using the Vanderbilt Prowler simulator. The MLP contributes to the WSN's lifespan extension, according to the simulation's findings.

- **Conditional privacy preserving security protocol for NFC applications**

AUTHORS: W. Kim, O.-R. Jeong, C. Kim, and J. So

THEME: NFC's integration with smart devices like credit cards has increased its adoption. The user's public key must be used during the key negotiation process at a predetermined value, according to the present implementation of NFC security protocols. The fixed components, such as the NFC public key, are where the message's relevance may be found. By combining the accompanying messages, a hacker can produce attributes depending on the user's public key. The profile that has been created could put users' privacy at risk. This study proposes conditional confidentiality security techniques based on pseudonyms to overcome these problems. The Protocol Data Unit for conditional privacy is further specified. By sending the conditionally private PDU via NFC terminals, users can let the other party know that they will interact using the framework suggested in this paper. With the use of NFC 1's physical properties, the suggested solution is able to reduce update costs and calculation overhead.

- **The dark side of the internet: Attacks, costs and responses**

AUTHORS: W. Kim, O.-R. Jeong, C. Kim, and J. So

THEME: When the Internet and Web technologies were initially created, it was believed that in a perfect society, everyone would act ethically. Unfortunately, the world is afflicted by the dark side that has emerged. Spam, malware, hacking, phishing, denial-of-service assaults, link theft, invasions of privacy, libel, fraud, and infringement of intellectual property rights are some examples of this. This paper examines several sorts of defences as well as the costs and reasons of attacks.

3. PROPOSED SYSTEM

This approach involves these steps:

- An open source, publicly accessible dataset called REFIT is utilised as a training set for research on end-use energy demand in the domestic building sector.
- The dataset was created from a research of smart home technology conducted in 20 homes in the UK. This study involved a number of sensor measurements, building evaluations, and household interviews.
- To improve machine performance, the inserted dataset is first examined for duplicates and null values. The dataset is then refined based on the total amount of energy used by the residence by ensuring that the recorded data values or vectors fall within the range of 0 and 1.
- The resulting dataset will be labelled with classes denoting "0" and "1" for no spam and spam, respectively. Afterwards, the dataset is splitted into 2 sub-datasets: say "train dataset" and "test dataset".
- Although both of the datasets were actually subsets of the original dataset the training dataset will be used to train the machine learning model.
- After acquiring the values and learning the new meaningful patterns from the training dataset the machine undergoes testing phase with new unseen testing dataset.
- At the end the dataset is trained to boost its performance by using machine learning classification algorithms like random forest, support vector machine (SVM), logistic regression algorithms.
- The end user input data values can be tested through a frontend webpage interface to check whether the data is spam or Ham.

ADVANTAGES

- Cost efficient and the web interface makes it easy to use for end user.
- The mapping of the user input vector with the specific class label will be more fast as the model is trained with efficient classifiers.
- Regardless of the criteria for climatic change, good efficiency and greater accuracy for large datasets.
- Our project, spam detection, is capable of filtering spam parameters that affect the IOT network without taking into account domain names or any other factors.
- In order to conserve energy and increase the lifespan of IoT systems, machine learning approaches aid in the development of protocols for light-weight access control.

ARCHITECTURE

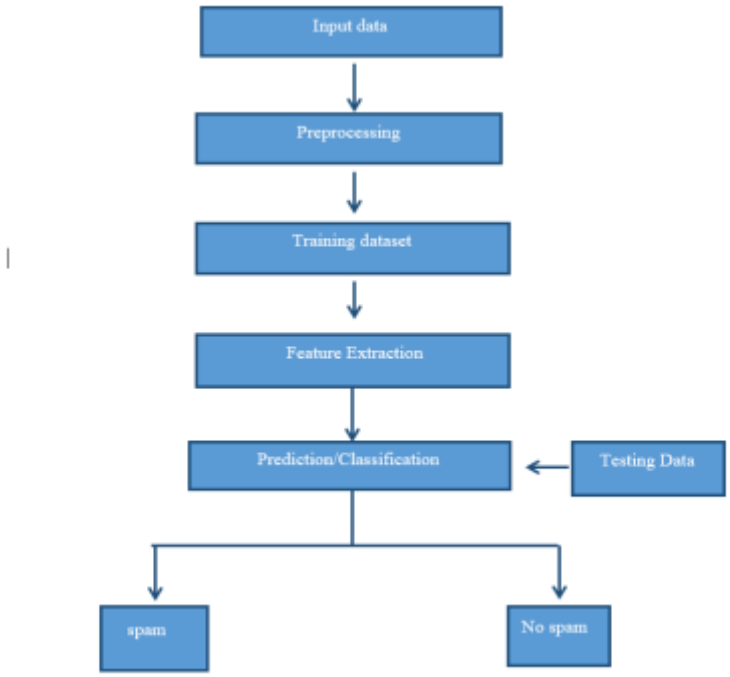

Fig.1 Architecture Diagram

HARDWARE REQUIREMENTS

- **System:** i3 Processor
- **Hard Disk:** 500 GB
- **Input Devices:** Keyboard, Mouse
- **RAM:** 2GB

SOFTWARE REQUIREMENTS

- **Operating System**: Windows 10/8/10 or MacOS 10.12+ or Ubuntu 14/16/18
- **IDE**: jupyter Notebook.
- **Framework:** Flask.
- **Libraries**: Numpy, scikit-learn-0.23.1, pandas, pickle, itertools.

4. MODULES

The framework usually consists of two modules namely system module, user module.

1. SYSTEM MODULE:

Data collection:

The actual process of creating a machine learning model and accumulating data starts with this stage. This stage is critical since the amount and quality of data we can gather will determine how effectively the model performs.

Import Dataset: The smart home dataset REFIT which incorporates nearly 503910 records will be imported.

REFIT Smart Home dataset Link: https://www.refitsmarthomes.org/datasets/

Data preparation:

We'll alter the data by removing any blanks and a few columns. We will start by listing the column names that we want to keep or retain. The remaining columns are dropped or removed after that, leaving only the ones we want to maintain. Last but not least, we remove the rows from the data set that have missing values. The dataset is refined on the basis of total energy consumption made by the house by making sure the recorded data values/vectors falling in the range between 0 and 1. In this data set we are taken 22 columns and 6980 rows after data pre-processing.

Model selection:

Two data sets are required when building a machine learning model: one for training and the other for testing. When creating a machine learning model, two datasets are needed: one for training and the other for testing. Let's divide this in two using a 70:30 split. The data frame will also be split into a feature column and a label column. We imported the sklearn train test split function here. Use it to partition the dataset after that.

Analyse and prediction:

The dataset will be given to our machine learning model, which will determine before relying on the IOT data whether the data contains spam parameters that can affect the IOT network trustworthiness and generate the accuracy and precision rates, which will vary depending on the model and algorithm chosen since the logistic regression, knn, support vector machine, and decision tree facilitate to classification problems they will be used to boost the model's accuracy and precision rates.

Saving trained model:

The first step is to store your trained and tested model into a.h5 or.pkl file using a library like pickle after you are comfortable utilising it in a production-ready environment. verify that Pickle is set up in your environment. The model will now be imported into the module and dumped as a.pkl file.

2. USER MODULE:

The system will be trained to boost its performance using classifiers namely random forest, logistic regression, support vector machine.

Offer input vectors: The user will volunteer input KW vectors that is not one of the features from the collection.

View Results: The merit of that concerned set of vectors will be revealed at the conclusion of the process.

5. ALGORITHM

Logistic regression algorithm:

Machine learning uses the categorization method known as logistic regression. The dependant variable is modelled using a logistic function. Due to the dichotomous structure of the dependant variable, there are only two viable classes (eg: either the data is spam or not). Hence, while working with binary data, this technique is employed when the data has a binary output, such as when it belongs to one class or another or is either a 0 or 1. Binary, multinomial, and ordinal logistic regression are the three main subtypes.

SVM algorithm:

A particular variety of classification algorithm based on margin maximisation is called a support vector machine (SVM). They minimise structural risks in order to increase the classifier's complexity and achieve great generalisation performance.

Random forest algorithm:

The algorithm's foundation is a decision tree, which significantly boosts accuracy. Simple decision trees are repeatedly created by Random Forest, which then uses the "majority vote" method to decide which label to return.

6. SYSTEM DESIGN

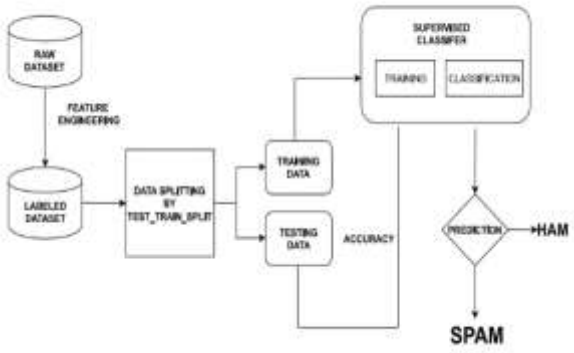

Fig 2 system design

The input dataset will undergo data pre-processing to eliminate the noisy, irrelevant data which finally results in clean labelled dataset. Then the resultant labelled dataset will be divided into training and testing dataset in equal proportion. In comparison to the testing dataset, the size of the training dataset is typically bigger. The ratios used to divide the train and test datasets are typically 80:20,60:40,70:30 and 90:10. We splitted in the proportion of 70:30. Although both of the datasets were actually subsets of the original dataset the training dataset will be used to train the machine learning model. After acquiring the values and learning the new meaningful patterns from the training dataset the machine undergoes testing phase with new unseen testing dataset to evaluate the accuracy of the trained model on new data. Finally the model is ready to predict the user input data values either spam or no spam.

7. CONCLUSION

IoT data must be analysed in order to be useful, but it is difficult to manually review the enormous amounts of data that IoT devices produce. Hence, a crucial part of most IoT solutions is automated analytics. Analytics tools are used on the telemetry data to generate descriptive reports, show data in dashboards and data visualisations, and trigger alerts and actions. The provided methodology will help determine whether the data gathered by IOT devices is spam or useful information. Since most IOT solutions will rely on sensed data produced by IOT devices, protecting the data's integrity and authenticity is crucial to ensuring the accuracy of these solutions. Its consideration will be crucial in the future year.

8. RESULT

REFERENCES

1. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram,"Blockchain for iot security and privacy: The case study of a smart home," in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, 2017, pp. 618–623.
2. E. Bertino and N. Islam, "Botnets and internet of things security," Computer, no. 2, pp. 76–79, 2017.
3. F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," Soft Computing, vol. 20, no. 1, pp. 343–357, 2016.
4. Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, "A system for denial-of-service attack detection based on multivariate correlation analysis," IEEE transactions on parallel and distributed systems, vol. 25,no. 2, pp. 447–456, 2013.
5. H. Eun, H. Lee, and H. Oh, "Conditional privacy preserving security protocol for nfc applications," IEEE Transactions on Consumer Electronics, vol. 59, no. 1, pp. 153–160, 2013
6. R. V. Kulkarni and G. K. Venayagamoorthy, "Neural network based secure media access control protocol for wireless sensor networks," in 2009 International Joint Conference on Neural Networks. IEEE, 2009, pp. 1680–1687.
7. z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, "Iot security: ongoing challenges and research opportunities," in 2014 IEEE 7th international conference on service-oriented computing and applications. IEEE, 2014, pp. 230–234.
8. C. Zhang and R. Green, "Communication security in internet of thing: preventive measure and avoid ddos attack over iot network," in Proceedings of the 18th Symposium on Communications & Networking. Society for Computer Simulation International, 2015, pp. 8–15.
9. W. Kim, O.-R. Jeong, C. Kim, and J. So, "The dark side of the internet: Attacks, costs and responses," Information systems, vol. 36, no. 3, pp. 675–705, 2011.
10. M. A A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2015.