IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Spam Detection Technique Using Machine Learning With Principle Component Analysis

Mrs. P. Immaculate Rexi Jenifer[1], Abinaya S[2], Banu Rithika.R[3], Madhu Bala. R[4]
Assistant Professor, Department of Science and Computer Engineering[1]
Students, Department of Science and Computer Engineering[2,3,4]
Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Tiruvarur, Tamil Nadu, India

Abstract: *A collection of millions of devices with sensors and actuators that are linked via wired or wireless channels for data transmission. Over the last decade, it has grown rapidly, with more than 25 billion devices expected to be connected by 2020. The amount of data released by these devices will multiply many times over in the coming years. In addition to increased volume, the device generates a large amount of data in a variety of modalities with varying data quality defined by its speed in terms of time and position dependency. In such an environment, machine learning algorithms can play an important role in ensuring biotechnology-based security and authorization, as well as anomalous detection to improve usability and security.On the other hand, attackers frequently use learning algorithms to exploit system vulnerabilities. As a result of these considerations, we propose that the security of devices be improved by employing machine learning to detect spam. Spam Detection Using Machine Learning Framework is proposed to attain this goal. Four machine learning models are assessed using multiple metrics and a vast collection of input feature sets in this framework. Each model calculates a spam score based on the input attributes that have been adjusted. This score represents the device's trustworthiness based on a variety of factors. In comparison to other current systems, the findings collected demonstrate the effectiveness of the proposed method.*

Keywords: Collection of data, Authorization, Anomalous detection, Support Vector Machine, K-nearest neighbour, Spam.

I. INTRODUCTION

Information exchange has become extremely simple and quick in the age of information technology. Users can exchange information on a variety of platforms from anywhere in the world. Email is the easiest, most cost-effective, and fastest method of transmitting information in the world. Emails, on the other hand, are vulnerable to a variety of attacks, the most popular and destructive of which is spam [1]. No one likes to receive emails that are irrelevant to their interests since they waste the time and resources of the recipients. Furthermore, dangerous content may be disguised in the form of attachments or URLs in these emails, resulting in security breaches on the host system [2].Spam is any irrelevant or unwanted message or email sent by an attacker to a large number of recipients via email or any other information sharing media [3]. As a result, there is a high demand for email system security. Viruses, rats, and Trojans may be contained in spam emails. This method is commonly used by attackers to entice consumers to use internet services. They may send spam emails with multiple-file attachments and packed URLs that direct users to harmful and spamming websites, resulting in data theft, financial fraud, and identity theft [4, 5].

Many email providers allow users to create keyword-based filters that filter emails automatically. This strategy, however, is ineffective since it is difficult, and users do not want to personalise their emails, which allows spammers to attack their accounts. The Internet of Things (IoT) has quickly become a part of modern life during the last few decades. The Internet of Things (IoT) has become a critical component of smart cities. There are numerous IoT-based social media sites and applications available. Spamming issues are on the rise as a result of the Internet of Things, according to Hindawi Security and Communication Networks Volume 2022, Article ID 1862888, 19 pages https://doi.org/10.1155/2022/1862888. (e scientists proposed a number of spam detection algorithms for detecting and filtering spam and spammers.) There are two sorts of existing spam detection methods: behaviour pattern-based approaches and semantic pattern-based approaches.

((each of these methods has its own set of limits and disadvantages.) Along with the expansion of the Internet and global communication, there has been a considerable increase in spam email [6]. Spam may be sent from anywhere in the world

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

thanks to the Internet, which hides the identity of the sender. Despite the fact that there are numerous antispam tools and approaches available, the spam rate remains high. Harmful emails with links to malicious websites that can harm a victim's data are among the most dangerous spams. Spam emails can also cause server response times to slow down by filling up memory and capacity. Every firm carefully assesses the available solutions to combat spam in their environment to accurately detect spam emails and avoid escalating email spam issues.

Whitelist/Blacklist [7], mail header analysis, keyword checking, and other well-known mechanisms for identifying and analysing incoming emails for spam detection are just a few examples. According to social networking experts, 40 percent of social networking accounts are used for spam [8]. (e spammers use popular social networking applications to transmit concealed links in the text to pornographic or other product sites aimed to sell something from fraudulent accounts to certain segments, review pages, or fan pages. (e obnoxious emails sent to the same types of people or organisations have recurring themes. It is possible to increase the detection of these types of emails by looking into these highlights. We can classify emails into spam and nonspam using artificial intelligence (AI) [9].

(This approach is achievable because to feature extraction from the headers, subject, and body of the messages.) We can categorise this data into spam or ham after extracting it based on its characteristics. Learning-based classifiers [10] are now widely employed to detect spam.

II. SPAM MESSAGES

The definition of email spam is vague because everyone has an opinion about it. At the moment, everyone's attention is drawn to email spam. In most cases, email spam consists of one-off messages sent in bulk by people you don't know. (The term spam comes from a Monty Python sketch [23], in which a Hormel canned beef item had a lot of annoying emphases.) While the term "spam" is said to have originally been coined in 1978 to refer to unsolicited email, it exploded in popularity in the mid-1990s, as we began to become more widely known outside of academic and research circles [24]. The development expenditure trick is a well-known example, in which a client receives an email with an offer that should result in a prize.In today's technological age, the dodger/spammer tells a scenario in which the unlucky victim requires immediate financial assistance so that the fraudster can amass a much larger sum of money, which they would subsequently share. When the unfortunate victim completes the instalment, the fraudster will either make a profit or avoid communication.

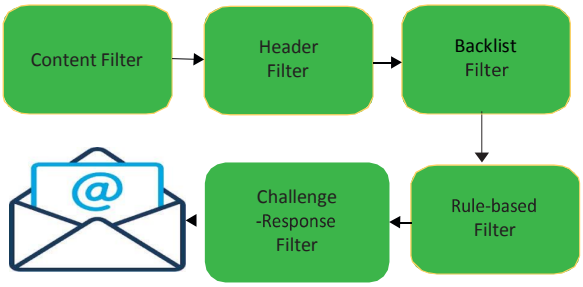


Fig 1 Block Diagram for Spam Messages

III. LITERATURE REVIEW

IoT systems, which include devices, services, and networks, are subject to network, physical, and application threats, as well as privacy breaches. These assaults are depicted in Figure 1. Let's take a look at some of the attacks that the attackers have launched.

- Denial of service (DDoS) attacks: To prevent IoT devices from accessing various services, attackers can flood the target database with unsolicited requests. Bots [3] are malicious queries generated by a network of Internet of Things devices. DDoS attacks might deplete all of the service provider's resources. It has the ability to block legitimate users and make network resources unavailable.
- RFID attacks: These are attacks against IoT devices at the physical layer. The device's integrity is compromised as a result of this assault. Attackers try to change data at the node level or while it is being transmitted via the network. At the sensor node, popular attacks include availability attacks, authenticity attacks, secrecy attacks, and cryptography key brute-forcing [4]. Password protection, data encryption, and restricted access control are

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

some of the remedies used to thwart such assaults.

- Attacks on the internet: The Internet-connected IoT gadget can access a variety of resources. Spammers utilise spamming strategies if they wish to steal information from other systems or if they want their target website to be viewed on a regular basis [5]. Ad fraud is a popular strategy used for this. For monetary gain, it generates artificial clicks on a particular website.Cyber criminals are a type of practising group.
- Near-field communication (NFC) assaults: These attacks mostly target electronic payment fraud. Unencrypted traffic, eavesdropping, and tag manipulation are all threats that could be launched.
- The conditional privacy protection provides a solution to this issue. Thus, using the user's public key, the attacker is unable to generate the identical profile [6]. The trustworthy service manager generates random public keys for this model. Various machine learning approaches have been widely employed to improve network security, including supervised learning, unsupervised learning, and reinforcement learning. Table I summarises the current machine learning techniques that aid in the identification of the aforementioned assaults. The next sections detail each machine learning technique in terms of its nature and role in attack detection.
- Supervised machine learning techniques: The models such as support vector machines (SVMs), random forest , naive Bayes, K-nearest neighbour (K-NN), and neural networks (NNs) are used for labelling the network for detection of attacks. In IoT devices, these models successfully detected the DoS, DDoS, intrusion and malware attacks [7] [8] [9] [10].
- Unsupervised machine learning techniques: In the absence of labels, these approaches outperform their equivalents [9]. It functions by creating clusters. Multivariate correlation analysis is used to detect DoS attacks in IoT devices [11].
- Reinforcement machine learning techniques: These types of models Allow an IoT system to use trial and error to choose security protocols and key settings in response to various assaults. Q-learning has been used to increase authentication performance and can also aid in virus detection. [12][9][13]

Machine learning approaches aid in the development of lightweight access control mechanisms that save energy and extend the life of IoT equipment. For example, the developed outside detection strategy uses K-NNs to address the problem of unregulated outer detection in WSNs [14]. The literature review explains how machine learning may be used to improve network security. As a result, multiple machine learning techniques are used to detect the presented problem of web spam in this research.

IV. PROPOSED SYSTEM

System Model

Smart devices are fully reliant on the digital world. The data obtained from these devices should be devoid of spam. Because data is collected from multiple domains, retrieving information from various IoT devices is a major difficulty.

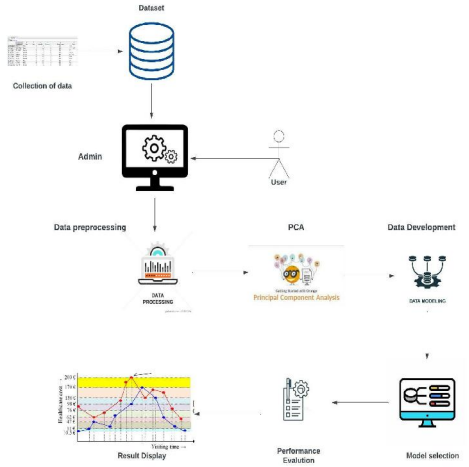


Fig 2 System Architecture for proposed system

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Because IoT involves various devices,a vast volume of data with heterogeneity and variation is generated.
This data is referred to as IoT data. Real-time, multi-source, rich, and sparse data are all characteristics of IoT data. Personal use is allowed, but republication and redistribution require IEEE approval. This paper has been accepted for publication in a future edition of this journal, but it is still being edited. Before the final publishing, the content may change.
Citation information:
If IoT data is stored, processed, and retrieved efficiently, its efficiency rises.
This proposal aims to reduce the occurrence of spam from these devices as defined by Eq.
1.min P(s) = א − ~s (1)
In Eq. 1, א refers to the collection of information. ~s is thevector of spam related information, which is subtracted fromא to decrease the probability of getting spam information from IoT devices.

V. PROPOSED METHODOLOGY

The online spam detection is addressed in this proposal to protect IoT devices from producing malicious in- formation. We looked at a number of machine learning strategies for detecting spam from IoT devices. The goal is to fix problems with IoT devices that are used in the home. However, the proposed methodology takes into account all aspects of data engineering before putting it to the test with machine learning models. The method that was employed to achieve the goal.

Feature Engineering: Machine learning algorithms perform correctly with the correct instances and attributes. We're all aware that the instances represent real-world data value gathered from real-world smart things placed around the planet. The feature engineering method is built around the extraction and selection of features.

Feature Reduction: This technique is used to reduce the number of dimensions in data. To put it another way, feature minimization. The approach for reducing the complexity of characteristics is described in Transactions on Industrial Informatics. Over-fitting, huge memory requirements, and compute power are all addressed by this technique. There are several methods for extracting features. The most often used is principal component analysis (PCA). However, in this suggestion, PCA is employed in conjunction with the IoT parameters. – Time for analysis: The dataset utilised in the studies involves data collected over an eighteen-month period. We used one month's worth of data for better results and accuracy. Because the weather is such a significant factor in the operation of IoT devices, the month with the most changes was chosen.

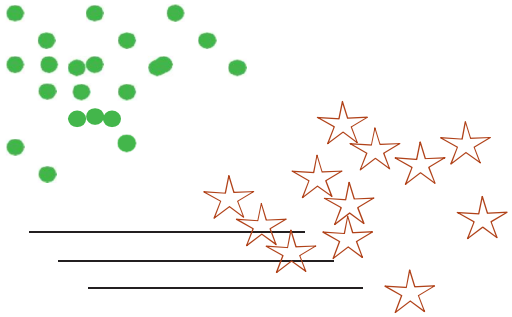
VI. METHODOLOGY

1. Data Collection
2. Data Preprocessing
3. Principle Component Analysis
4. Performance Evaluation

- **Data Collection:** The practise of acquiring and analysing data from a variety of sources is known as data collection. Data must be collected and kept in a form that makes sense for the business problem at hand in order to use it to develop viable artificial intelligence (AI) and machine learning solutions.
- **Data Pre-processing:** Machine learning validation approaches are used to calculate the error rate of the Machine Learning (ML) model, which is as close to the genuine error rate of the dataset as possible. Validation approaches may not be required if the data volume is large enough to be representative of the population. However, in real-world circumstances, it is necessary to work with data samples that are not always representative of the population of a dataset. Duplicate the value and the data type description to identify the missing value, whether it is a float variable or an integer variable. While tuning model hyper parameters, a sample of data is employed to offer an unbiased evaluation of a model fit on the training dataset.
- **Principle Component Analysis:** This procedure entails gathering preprocessed data in order to translate it into a model creation process. Over-fitting, huge memory requirements, and compute power are all addressed by this technique. There are several methods for extracting features. The most widely used method is principal component analysis (PCA). However, in this suggestion, PCA is employed in conjunction with the IoT parameters.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

- **Performance Evaluation:** As competence on the validation dataset is incorporated into the model setup, the evaluation becomes increasingly biassed. The validation set is used to test a model, although it is only used on a regular basis. This data is used by machine learning specialists to fine-tune the model hyper parameters. Data collection, analysis, and the process of addressing data content, quality, and organisation can be time-consuming.
- **Support Vector Machine:** The support vector machine (SVM) is a powerful and important machine learning model. SVM is a formally defined supervised learning classifier that uses labelled examples for training and outputs a hyperplane for classifying fresh data. Decision planes separate a group of objects belonging to multiple class memberships. Linear support vector machines' classification concept. Some circles and stars are referred to as objects in the diagram. These objects can be classified into one of two categories: stars or dots. The isolated lines define which things are green and which are brown. The items on the lower side of the plane are brown stars, while the objects on the upper side are all green dots, indicating that two unique objects are classified into two different classes. If the model is given a new black circle object, it will categorise it into one of the classes using the training examples provided during the training phase.



- **KNN:** The K-Nearest Neighbour algorithm is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms. The K-NN algorithm assumes that the new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories. The K-NN method stores all available data and classifies a new data point based on its similarity to the existing data. This means that new data can be quickly sorted into a well-defined category using the K-NN method. The K-NN algorithm can be used for both regression and classification, but it is more commonly utilised for classification tasks. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data. It's also known as a lazy learner algorithm since it doesn't learn from the training set right away; instead, it saves the dataset and performs an action on it when it comes time to classify it. During the training phase, the KNN algorithm simply stores the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

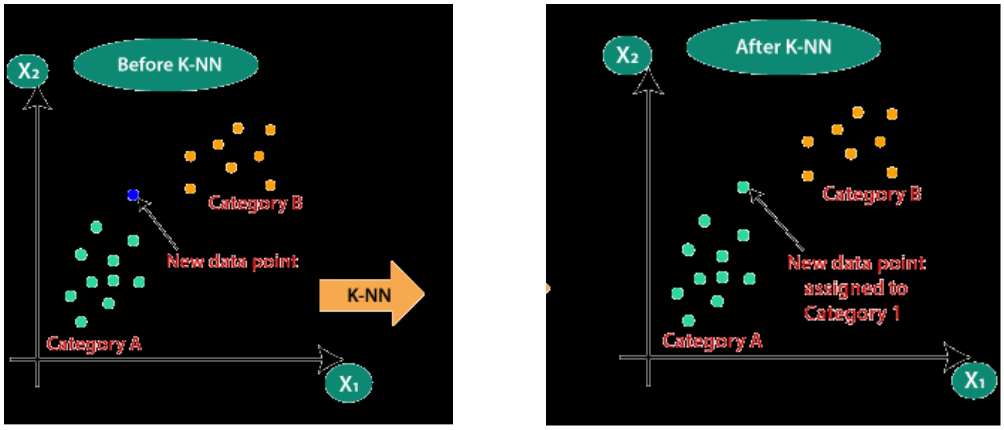


Fig 3 KNN Classifier Algorithm

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

VII. RESULT AND DISCUSSION

Following our discussion of the literature, we discovered that the vast majority of datasets used to train, test, and apply various models are created synthetically. There aren't enough examples for analysis, and categorising all of the supervised model data is difficult. As a result of the synthetic datasets used to train the models, the findings of the classifiers are not completely reliable. Because a large number of machine learning models are now utilised for email spam detection or filtering, these are not indicative of real-world spam reviews. The three learning algorithms presented here, logistic regression, Nave Bayes, and support vector machine (SVM), are extensively utilised and outperform the other algorithms in the majority of the research. In general, SVM provides the best results. It is frequently defeated by Nave Bayes and logistic regression. However, because it is not compared to all other algorithms, SVM should not be deemed the best. Several learning models based on different feature engineering techniques. By examining and monitoring a variety of strategies, this survey study elaborates on the existing machine learning-based spam filtering techniques and models. The findings are reviewed after a review of multiple spam filtering algorithms and a summary of the accuracy of several proposed approaches depending on various parameters. We conclude that all spam filtering methods are effective. Some should be assessed in future studies based on a variety of remarkable results, while others are attempting to improve accuracy through various approaches. Despite the fact that they are all effective, the spam filtering system still lacks some of the features that researchers are most concerned about. They're attempting to develop next-generation spam filtering systems that can handle multimedia data and effectively filter spam email.

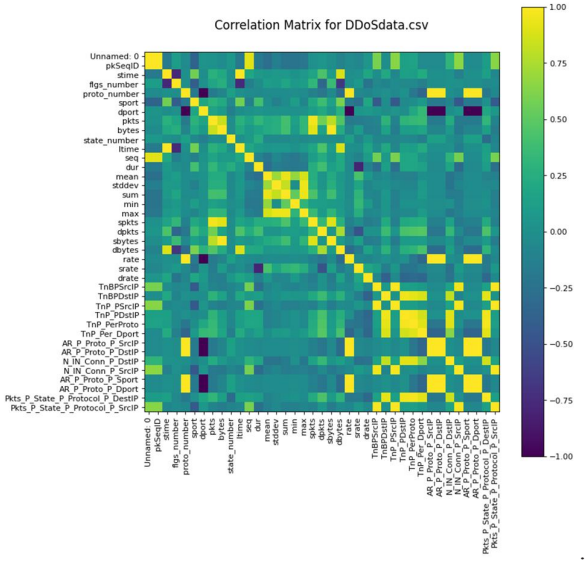


Fig 4 Correlation Matrix for proposed system

VIII. CONCLUSION

Using machine learning models, the proposed framework detects the spam properties of IoT devices. The feature engineering approach is used to pre-process the IoT dataset utilised in the tests. Each IoT appliance is given a spam score after experimenting with machine learning models in the framework. This clarifies the conditions that must be met for IoT devices to function properly in a smart home. In the future, we intend to take into account the climatic and environmental characteristics of IoT devices in order to make them more secure and reliable.

REFERENCES

[1]. Aaisha Makkar, Sahil (GE) Garg, Neeraj Kumar, M. Shamim Hossain, Ahmed Ghoneim, Mubarak Alrashoud,"An Efficient Spam Detection Technique for IoT Devices using Machine Learning" ,IEEE Transactions on Industrial Informatics (Volume: 17, Issue: 2, Feb. 2021)

[2]. Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, andS. 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.

[3]. A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for iot security and privacy: The case study of a smarthome," in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, 2017, pp. 618–623.

[4]. E. Bertino and N. Islam, "Botnets and internet of things security," Computer, no. 2, pp. 76–79, 2017.

[5]. C. Zhang and R. Green, "Communication security in internet of thing: preventive measure and avoid ddos attack over iot network," Proceedings of the 18th Symposium on Communications & Networking. Society for Computer Simulation International, 2015, pp. 8–15.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications,"IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996–2018, 2014.

[10]. 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.

[11]. 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.

[12]. N. Sutta, Z. Liu, and X. Zhang, "A study of machine learningalgorithms on email spam classification," in Proceedings of the 35th International Conference, ISC High Performance 2020, vol. 69, pp. 170–179,Frankfurt,Germa.

[13]. L. Xiao, Y. Li, X. Huang, and X. Du, "Cloud-based malware detection game for mobile devices with offloading," IEEE Transactions on Mobile Computing, vol. 16, no. 10, pp. 2742–2750, 2017.

[14]. J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," Knowledge and information systems, vol. 34, no. 1, pp. 23–54, 2013.

[15]. I. Jolliffe, Principal component analysis. Springer, 2011.

[16]. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar, pp. 1157–1182, 2003.

[17]. L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 856–863.

[18]. A. H. Sodhro, S. Pirbhulal, and V. H. C. de Albuquerque, "Artificial intelligence driven mechanism for edge computing based industrial applications," IEEE Transactions on Industrial Informatics, 2019.

[19]. A. H. Sodhro, Z. Luo, G. H. Sodhro, M. Muzamal, J. J. Rodrigues, and V. H. C. de Albuquerque, "Artificial intelligence based qos optimization for multimedia communication in iov systems," Future Generation Computer Systems, vol. 95, pp. 667–680, 2019.

[20]. L. University, "Refit smart home dataset," https://repository.lboro.ac.uk/ articles/REFIT Smart Home dataset/2070091, 2019 (accessed April 26, 2019).

[21]. R, "Rstudio," 2019 (accessed October 23, 2019

[22]. T. Vyas, P. Prajapati, and S. Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering," in Proceedings of the 2015 IEEE international conference on electrical, computer and communication technologies (ICECCT), IEEE, Tamil Nadu, India, March 2015.

[23]. L. N. Petersen, "(e ageing body in monty Python live (mostly)," European Journal of Cultural Studies, vol. 21, no. 3, pp. 382–394, 2018.

[24]. L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, and J. D. Tygar, "Characterizing botnets from email spam records," LEET, vol. 8, pp. 1–9, 2008.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

[25]. W. N. Gansterer, A. G. K. Janecek, and R. Neumayer, "Spam filtering based on latent semantic indexing," in Survey of Text Mining II, pp. 165–183, Springer, New York, NY, USA, 2008.

[26]. D. Lee, M. J. Lee, and B. J. Kim, "Deviation-based spamfiltering method via stochastic approach," EPL (Europhysics Letters), vol. 121, no. 6, Article ID 68004, 2018.

[27]. A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," Telecommunication Systems, vol. 68, no. 4, pp. 687–700, 2018.

[28]. M. F. N. K. Pathan and V. Kamble, "A review various techniques for content based spam filtering," Engineering and Technology, vol. 4, 2018.

[29]. A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," EURASIP Journal on Information Security, vol. 2016, no. 1, p. 9, 2016.

[30]. A. Bhowmick and S. M. Hazarika, "Machine learning for E-mail spam filtering: review, techniques and trends," 2016, https://www.researchgate.net/publication/303812063_Machine_Learning_for_ E-mail_Spam_Filtering_ReviewTechniques_and_Trends

[31]. M. Bassiouni, M. Ali, and E. A. El-Dahshan, "Ham and spam e-mails classification using machine learning techniques," Journal of Applied Security Research, vol. 13, no. 3, pp. 315– 331, 2018.

[32]. J. R. M´endez, T. R. Cotos-Yañez, and D. Ruano-Ord´as, "A new semantic-based feature selection method for spam filtering," Applied Soft Computing, vol. 76, pp. 89–104, 2019.

[33]. R. Alguliyev and S. Nazirova, "Two approaches on implementation of CBR and CRM technologies to the spam filtering problem," Journal of Information Security, vol. 3, no. 1, Article ID 16724, 2012.

[34]. E. Alpaydin, Introduction to Machine Learning, MIT Press, Cambridge, UK, 2020.

[35]. E. P. Sanz, J. M. Gomez Hidalgo, and J. C. Cortizo P ´ erez, ´ "Chapter 3 email spam filtering," Advances in Computers, vol. 74, pp. 45–114, 2008.

[36]. S. Pitchaimani, V. P. Kodaganallur, and C. Newell, "Systems and methods for controlling email access," Google Patents, 2020.

[37]. A. d. A. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, "Neural-symbolic computing: an effective methodology for principled integration of machine learning and reasoning," Journal of Applied Logic, vol. 6, 2019.

[38]. A. Singh, N. (akur, and A. Sharma, "A review of supervised machine learning algorithms," in Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, New Delhi, India, March 2016.

[39]. J. Tanha, M. van Someren, and H. Afsarmanesh, "Semi-supervised self-training for decision tree classifiers," International Journal of Machine Learning and Cybernetics, vol. 8, no. 1, pp. 355–370, 2017