International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

IJIEMR Transactions, online available on 22th Dec 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 12

10.48047/IJIEMR/V11/ISSUE 12/75

Title **An Email Spam Filtering Approach Using a Collaborative Reputation-Based Vector Space Model**

Volume 11, ISSUE 12, Pages: 606-614

Paper Authors **Masrath Parveen, Dr. Saurabh Pal, Dr. Venkateswara Rao CH**

International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

An Email Spam Filtering Approach Using a Collaborative Reputation-Based Vector Space Model

Masrath Parveen[1], Dr. Saurabh Pal[2], Dr. Venkateswara Rao CH[3]

[1]Research Scholar, Dept of CSE, V.B.S.Purvanchal University, Jaunpur
[2] Department of CSE, V.B.S.Purvanchal University, Jaunpur
[3]Department of CSE, Siddhartha Institute of Engineering and Technology, Hyderabad

Abstract – We suggest a novel Collaborative Reputation-based Vector Space Model (CRVSM) in this paper for the identification of spam email. across order to detect spam emails across a wide area, CRVSM employs a vector space model to describe the feature vectors in multidimensional vector space. To speed up email spam detection, we group the emails into five groups. With a maximum and lowest threshold range, we compute the maximum similarity measure to lower the amount of false positives and false negatives. In addition, we employ a reputation evaluation tool that assesses the reporter's level of credibility when verifying the email as spam or not. In terms of email spam detection, the CRVSM technique has good efficiency and good results. In terms of email spam detection, the CRVSM technique has good efficiency and good results. Utilising measures like false positive rate, false negative rate, detection accuracy, and detection time, the performance of the CRVSM model has been assessed. The performance results unmistakably demonstrate that CRVSM surpasses the currently used detection algorithms and effectively classifies emails that arrive as spam or non-spam with lower FPR and FNR values.

Keywords: Feature, Cluster, Collaborative, Vector, Spam Email, Similarity

1. Introduction

Email is one of the widely used computer-mediated communication methods, and because sending and receiving emails has no cost, it has replaced other means of contact for many individuals. The most well-known type of malware attack is spam email. 29 Emails are typically sent using the Multipurpose Internet Mail Extension (MIME) standard and the Simple Mail Transfer Protocol (SMTP) protocol [1], making them more susceptible to virus assaults. According to studies, spam makes up more than 85% of contemporary email traffic [2].

Email spam uses up resources on computers, such as storage space, network bandwidth, processing power, and traffic abuse. Email identification for spam is increasingly crucial in a personalized and social setting since spam emails cost the online community a great deal of money every year. Spammers are constantly coming up with new techniques to get over filters, and new solutions are being developed in response to keep spam emails hidden from users. Researchers have so far developed a number of methods to combat email spam, however none of these methods is a perfect answer [3].

A Collaborative Reputation-based Vector Space Model (CRVSM) [4] for email spam detection is presented in this paper. In order to recognise spam emails in a vast area, this approach identifies all incoming emails as vectors and clusters them into five groups to speed up email spam detection. The CRVSM model uses maximum similarities threshold values while performing similarity tests, which helps to decrease the amount of false positives and false negatives and improve detection accuracy. Additionally, the CRVSM model does reputation evaluation to ascertain each reporter's level of trust whether classifying an email as spam or not. The Map Reduce tools used by the dynamic CRVSM model, which leverages big data analytics, increase the model's efficiency over time.

2. Literature review

Email, often known as electronic mail, is frequently used abusively. Spam emails are a strain on mailing systems because they waste millions of subscribers' priceless time and resources. 71.9% of email traffic is spam, according to the Symantec Intelligence Review (Symantec Intelligence Report, 2013) [5]. Numerous studies on the identification of email spam have been conducted. According to the current study, spam is referred to be "unsolicited commercial email messages." We divide the email spam methods for identification into two groups based on the currently available research: statistical approach and rule-based approach [1,6].

Three-way email spam separating reduces the possibility of misclassification by grouping incoming emails into three different folders: spam, legitimate, and suspicious [1]. The spam folder contains emails that we believe are spam, the legitimate folder (inbox) contains emails that we suppose are legitimate, and the suspicious folder contains emails about which we are unsure based on the information at hand [7]. In order to determine three-way classifications when numerous criteria are present from a tradeoff perspective, game-theoretic rough sets (GTRS) [5] have recently been presented. The design of the game and the GTRS's repetition learning mechanism are thoroughly investigated [8]. The UCI Spambase Data Set is used as the basis for the experiment [9]. The experimental finding demonstrates that by somewhat sacrificing precision, the three-way filters developed by GTRS may greatly increase coverage [10].

The email properties are statistically described in the statistical method, and they are automatically categorised using machine learning algorithms like the Bayesian filter, naive Bayesian classifiers [11], Decision trees [12], the maximum entropy model [13], memory based learning [14], support vector machines[15], and k-nearest neighbor classifier [16] are among the techniques discussed in this article and enhancing. To examine the classification criteria of these methods, a set of pre-classified email messages, commonly referred to as training samples, is needed [17].

Spam email growth contributes to traffic jams, decreased productivity, and phishing, which has become a major issue for our society. Additionally, every year there are more spam emails than ever [18]. Thus, spam e-mail filtering is a significant, meaningful, and difficult subject. Finding a practical way to filter potential spam emails is the goal of this research [19]. The suggested method just uses the content of emails to create keyword corpora, along with some text processing to deal with obfuscation strategy. The CSDMC2010 SPAM corpus dataset [20], which included 4292 emails in the testing dataset and 4327 emails in the training dataset, was used to evaluate the method. The experimental findings demonstrate the proposed algorithm's accuracy of 92.8% [21].

Spam emails violate private information and are pricy forms of unsolicited mail. Spam email encroachment chases users and wastes network resources. To far, a number of filtering techniques have been used, most of them are based on various Machine Learning algorithms [22]. The accuracy of some of these techniques varies, and a few of them are expensive in terms of computer complexity. The suggested method uses decision tree algorithms for email filtering, which are straightforward and provide superior accuracy [23].

For the purpose of detecting email spam, a hybrid model combining differential evolution (DE) and the negative selection algorithm (NSA) was presented [12]. Hazardous URLs in Email have been found using a limited feature set method. a framework for spam filtering in a cloud context that is scalable. As decisions made by one person or one system are unreliable and must be verified, autonomous decision-making on Email detection of spam becomes less effective [24]. Therefore, group decision-making provides the greatest results for email spam detection. This approach necessitates the collaboration of a collection of receivers regarding their classification of an email as spam or not. The most common cooperative email spam filters are Pizor, Cloudmark, Vipul's Razor, and Distributed Checksum Clearinghouse [25].

The majority of current methods are implemented at the server, which places a pressure on the server to manage such massive activities in a short amount of time. In order to shorten the detection time to a larger extent, it would be beneficial if the anti-spam procedures were deployed cooperatively at the receiver side. Taking all of these factors into consideration, we therefore propose CRVSM, a novel email spam detection method that first operates feature extraction through a vector space model, then similarity detection, and finally a collaborative reputation procedure for verifying the reputation result.

3. Proposed model to perform defense

In a timely manner, the “Collaborative Reputation-based Vector Space Model (CRVSM)” discovers and reduces spam emails. Figure 1 depicts the CRVSM model's operational model. To successfully conduct the defence, the CRVSM model consists of three separate phases: Feature Extraction, Similarity Detection, and Collaborative Reputation Evaluation.

While the feature extraction step, spam emails are accurately filtered while all potential features are recovered from the received email using a vector space model. A soft cosine measure of similarity in the vector space is used to compare the related spam email, which was discovered during the feature extraction step, and another new email that was recently received.

3.1 The Feature Extraction process Using Vector Space Model

The process of extracting useful, non-redundant features into a new feature space from a set of data points measured is known as feature extraction. The crucial step in email spam detection approaches is feature extraction because it uses fewer resources while explaining vast data and yields reliable findings. The step-by-step process of the CRVSM protocol’s feature extraction 608tilizing a vector space model is explained below.

Step 1: Applying an external stop-word list, the feature extraction step first extracts all of the email’s crucial features and stopspellings. Stop-words such (‘is’, ‘a’, ‘an’, and ‘the’) are eliminated from the dataset to make it easier for characteristics to be represented as vectors by the VSM.

Step 2: In this stage, features are grouped into five distinct clusters, including cluster 1 for email headers, cluster 2 for attachment files, cluster 3 for text-only emails, cluster 4 for emails with text-and-html content, and cluster 5 for embedded resources. Emails are clustered for two reasons: to speed up searches and enable concurrent feature gain computation across clusters. Each email contains clusters and is shown as: $e_j = \{C_1, C_2, \ldots, C_n\}$

Step 3: is to represent the email as a multidimensional "vector space model (VSM)". The 4-tuple "Vector Space Model" is represented as: [E, C, $Fun_{(m,e)}$, R], where e denotes emails ranging from e_1 to e_n where each of which comprises of l features in a C group of clusters denoted in an email through R as a evolution function to determine the trust value through a reporter. And $Fun_{(m,e)}$ will perform similarity identification to measure email spam similarity score with the assumption that, emails are ordered alphabetically in terms that can be seen in the multidimensional vector space.

$$tf_{i,k,j} = \frac{b_{i,k,j}}{\sum_{K} b_{i,k,j}} \tag{1}$$

Step 4: Explain the meaning of the terms "frequency" and "inverse document frequency values," which are used to determine feature gain where frequency is denoted as: $tf_{i,k,j}$ and the frequency is measured as denoted in eq.(1) where $b_{i,k,j}$ is possible number of occurrences in each term of b_i over each cluster C_k over j^{th} email and $\sum$ determines the summation of all terms that are present over a specific email.

Step 5: Calculate the Feature Gain FG(l) for each of the features l in email cluster k, which determines the spam ratio of the possibility that a feature is spam and is determined in equation 2.

$$\mathrm{FG(l)} = \sum_{\mathrm{f_l} \in \mathrm{kB_j}} \sum_{\mathrm{C_i}} \mathrm{P(f_i, C_i)} * \frac{\mathrm{P(f_l, C_i)}}{\mathrm{P(f_1)} * \mathrm{P(C_i)}} \tag{2}$$

Where fl denotes l^{th} feature and $P(f_l)$ is the probability of features related to C_i dataset over the class comprising of {1,2} set denoted by $P(C_i)$ with the spam ratio in the proposed architecture shown in figure 1.

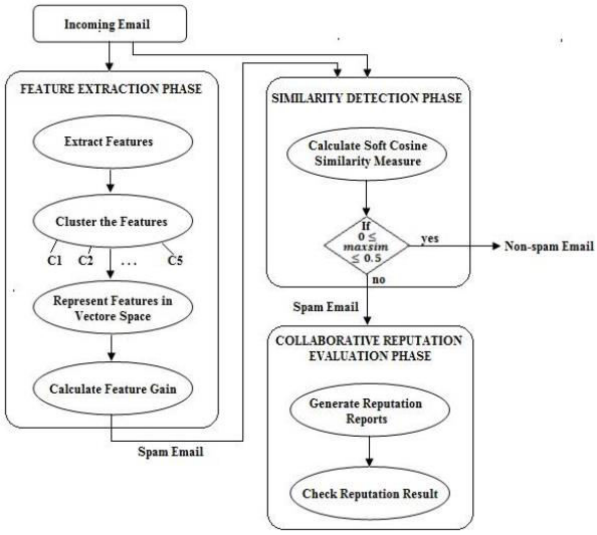

Fig.1.Architecture of the proposed system

All the five steps are implemented as shown in figure 1 sequence. Various reporters assess the accuracy of the spam emails that were discovered during the feature extraction and similarity detection phases. This calls for the computation of a reputation function that, according to Equation (2), creates reputation reports.

Reports are created by various reporters and given to the appropriate reporters of that email for review using the reputation function. This stage has been employed as the last evaluation stage to identify and validate a spam email. Every incoming email is subject to a reputation function computation, which determines the reporter's trust in an email based on feature gain.

4. Experimental Results

The suggested CRVSM protocol has been implemented and evaluated using the Java programming model. 1.4 million emails were included in the dataset that was used for the experiments.

A 0.5% spam ratio was used to analyse the performance of the CRVSM protocol with 5000, 10000, 15000, 20000, and 25000 emails. On a PC with a general-purpose processor (Intel(R) i5 processor) running at 2.67 GHz and 8 GB of RAM, the CRVSM protocol has been used. “False Positive Rate, False Negative Rate, Detection Accuracy, Spam Detection Rate, Spam Detection Time, Network Service Ratio, and Overall Throughput” were the metrics utilised to assess CRVSM's performance. Additionally, the efficacy of CRVSM has been evaluated in comparison to other methods, such as the “Fradulent Email Detection Model (FEDM)”. It has been demonstrated that CRVSM outperforms FEDM, MLP, VSM, and PM in terms of False Positive Rate (FPR), False Negative Rate (FNR), Detection Accuracy, Spam Detection Rate, Spam Detection Time, Network Service Ratio, and Overall Throughput.

By altering the percent of collaborative reporters, the total number of senders, and the quantity of incoming emails for each sender, the performance metrics of the CRVSM protocol have been assessed. The proportion of cooperative reporters has ranged from 20% to 80% of the overall recipients of pertinent correspondence. The reporters cooperate well when working together. Between 100 and 500 people have sent emails, and between 5000 and 25000 people have received them. The CRVSM model is dynamic and becomes better with use.

To identify fraudulent emails, the FEDM model employs a variety of classification algorithms, including Support Vector Machine, Naive Bayes, J48, Cluster-based Classification, and sophisticated feature selection methods. To improve accuracy, FEDM just looks at the email contents, such as the body and topic. The email features are represented by FEDM using a vector space model.

In order to identify spam emails, MLP, or a Multilayer Perceptron model, employs rule-based filtering and machine learning techniques. To improve accuracy, MLP looks at the email's body, subject, email header, and specific keywords.

A network level-based filter is used by PM, a pipeline-based model, to identify spam emails. Filters used to function in various stages are arranged in a pipeline order and include DNS blacklists, SYN packet filters, traffic behaviour filters, and message content filters. To achieve high detection accuracy, emails must flow through these filters in a pipeline format.

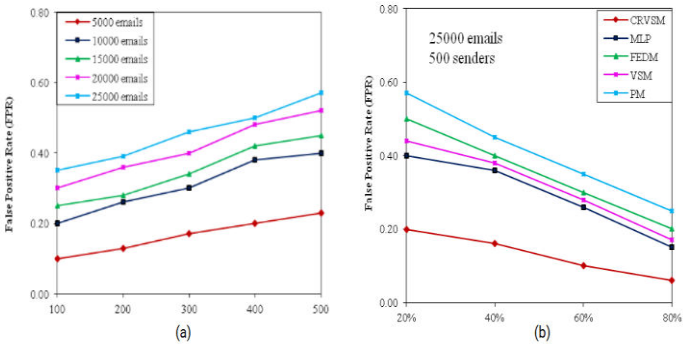

Fig.2. (a) FPR for CRVSM Vs Number of Email Senders (b) Comparison of FPR for CRVSM

The experiment is carried out to examine the FPR for 100 to 500 senders by sending 5000, 10000, 15000, 20000, and 25000 emails, respectively. Figure 2(a) displays the FPR for various email sender counts and email counts. The number of false positives has been found to decrease when the number of senders grows with rising number of emails since it has been found that with fewer senders, only fewer emails reach the recipient. The FPR value rises when there have more false positives, and vice versa.

In an experiment, the FPR of the CRVSM protocol is compared to those of four other protocols, namely MLP, FEDM, PM, and VSM, with different percentages of collaborative reporters. The experiment involves 25000 emails and 500 senders.

The analysis of FPR for CRVSM with four standard techniques for various percentages of collaborative reporters is shown in Figure 2(b). According to observations, the FPR of CRVSM lowers as the total amount of collaborative reporters rises, producing outcomes that are superior to those of traditional procedures. According to the data, CRVSM generates an FPR of 0.20 with 500 senders and 25000 emails sent, and for 20% of collaboration reporters. However, as the number of collaborative reporters rises (to 80%), the FPR of CRVSM drops to 0.06 and delivers better outcomes than existing protocols. By producing less FPR than the other four methods, CRVSM performs better.

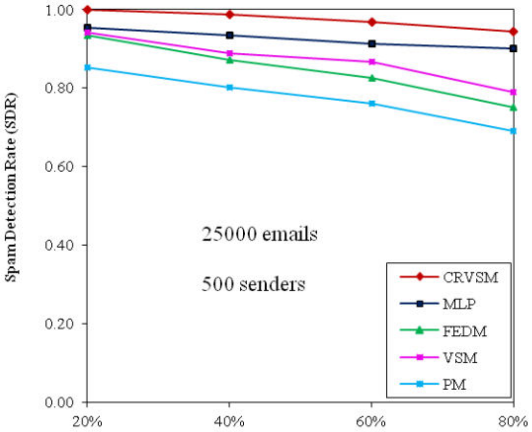

Fig.3. Comparison of Spam Detection Rate for CRVSM

For increasing percentages of collaborative reporters, the experiment compares the Spam Detection Rate of the CRVSM protocol with MLP, FEDM, VSM, and PM. Figure 3 compares the SDR for CRVSM with other procedures for various collaborative reporter percentages. It has been noted that the CRVSM protocol operates better than MLP, FEDM, VSM, and PM due to its high spam detection rate. This is due to the fact that the percentage of collaborative reporters, the diversity of threshold values chosen for determining the maximum similarity, and the computation of feature gain all affect the spam detection rate of CRVSM. As a result, CRVSM achieves an average 0.98 spam detection rate.

5. Conclusion

This study suggests an email spam filtering approach using a collaborative reputation-based vector space model where the experimental findings and performance findings of the innovative CRVSM protocol. According to the experimental findings, the unique CRVSM protocol performs better than previous protocols including “MLP, FEDM, VSM, and PM”. By lowering the false positive and false negative rate and so raising detection accuracy to a higher extent, the CRVSM protocol successfully detects spam emails accurately. By shortening the spam detection time, the CRVSM protocol enables timely detection. The CRVSM protocol further demonstrates its effectiveness by having a high rate of spam detection. The CRVSM protocol achieves an acceptable network service ratio and total throughput, guaranteeing system performance.

References

[1] P. Liu and T. -S. Moh, "Content Based Spam E-mail Filtering," 2016 International Conference on Collaboration Technologies and Systems (CTS), Orlando, FL, USA, 2016, pp. 218-224, doi: 10.1109/CTS.2016.0052. M. A. Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, “Laptop Price Prediction using Machine Learning Algorithms”, 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.

[2] Mohammed Ali Shaik, Praveen Pappula, T Sampath Kumar,

"Predicting Hypothyroid Disease using Ensemble Models through Machine Learning Approach", European Journal of Molecular & Clinical Medicine, 2022, Volume 9, Issue 7, Pages 6738-6745. https://ejmcm.com/article_21010.html

[3] M. A. Shaik, S. k. Koppula, M. Rafiuddin and B. S. Preethi, (2022), "COVID-19 Detector Using Deep Learning", International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 443-449, doi: 10.1109/ICAAIC53929.2022.9792694.

[4] A. Subasi, S. Alzahrani, A. Aljuhani and M. Aljedani, "Comparison of Decision Tree Algorithms for Spam E-mail Filtering," 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2018, pp. 1-5, doi: 10.1109/CAIS.2018.8442016.

[5] Mohammed Ali Shaik and Dhanraj Verma, (2022), "Prediction of Heart Disease using Swarm Intelligence based Machine Learning Algorithms", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020025-1–020025-9; https://doi.org/10.1063/5.0081719, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020025-1 to 020025-9.

[6] M. A. Shaik and Dhanraj Verma, (2022), "Predicting Present Day Mobile Phone Sales using Time Series based Hybrid Prediction Model", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020073-1–020073-9; https://doi.org/10.1063/5.0081722, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020073-1 to 020073-9

[7] in Srinidhi, Jia Yan & Giri Kumar Tayi 2015, ‘Allocation of resources to cyber-security: The effect of misalignment of interest between managers and investors’, Decision Support Systems, July 2015, http://dx.doi.org/10.1016/j.dss.2015.04.011, vol. 75, pp. 49-62.

[8] Mohammed Ali Shaik, MD.Riyaz Ahmed, M. Sai Ram and G. Ranadheer Reddy, (2022), "Imposing Security in the Video Surveillance", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020012-1–020012-8; https://doi.org/10.1063/5.0081720, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020012-1 to 020012-8.

[9] M. A. Shaik, Geetha Manoharan, B Prashanth, NuneAkhil, Anumandla Akash and Thudi Raja Shekhar Reddy, (2022), "Prediction of Crop Yield using Machine Learning", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020072-1–020072-8; https://doi.org/10.1063/5.0081726, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020072-1 to 020072-8.

[10] Mohammed Ali Shaik, Dhanraj Verma, (2021), Agent-

MB-DivClues: Multi Agent Mean based Divisive Clustering, Ilkogretim Online - Elementary Education, Vol 20(5), pp. 5597-5603, doi:10.17051/ilkonline.2021.05.629

[11] Byrnea, DJ, David Morganb, Kymie Tana, Bryan Johnsona & Chris Dorrosa 2014, 'Cyber Defense of Space-Based Assets: Verifying and Validating Defensive Designs and Implementations', Conference on Systems Engineering Research (CSER 2014), Science Direct, vol. 28, pp. 522-530.

[12] Mohammed Ali Shaik and Dhanraj Verma, (2020), Enhanced ANN training model to smooth and time series forecast, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022038, doi.org/10.1088/1757-899X/981/2/022038

[13] M. A. Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027

[14] Ashish Malviya, Glenn A Fink, Landon Sego & Barbara Endicott-Popovsky 2011, 'Situational Awareness as a Measure of Performance in Cyber Security Collaborative Work', Eighth International Conference on Information Technology: New Generations, pp. 937-942.

[15] Mohammed Ali Shaik and Dhanraj Verma, (2020), Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study, 2020 IOP Conf. Ser.:Mater.Sci.Eng. 981 022041, doi.org/10.1088/1757-899X/981/2/022041

[16] Mohammed Ali Shaik, "Time Series Forecasting using Vector quantization", International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238,Volume-29,Issue-4 (2020), Pp.169-175.

[17] Arunabha Mukhopadhyay, Samir Chatterjee, Debashis Saha, Ambuj Mahanti & Samir K Sadhukhan 2013, 'Cyber-risk decision models: To insure IT or not?', Decision Support Systems, http://dx.doi.org/ 10.1016/j.dss.2013.04.004, Volume 56, December 2013, pp. 11-26.

[18] Mohammed Ali Shaik, "A Survey on Text Classification methods through Machine Learning Methods", International Journalof Control and Automation (IJCA), ISSN:2005-4297,Volume-12,Issue-6 (2019), Pp.390-396.

[19] Andreas GK Janecek, Wilfried N Gansterer & Ashwin Kumar, K 2008, 'Multi-Level Reputation-Based Greylisting', in proc. of Third International Conference on Availability, Reliability and Security ARES 08, 4-7 March 2008, Barcelona, Spain.

[20] Mohammed Ali Shaik, P. Praveen, T. Sampath Kumar, "Integration and application of Fog, IoT and Edge Computing", Fog Computing: Concepts, Frameworks, and Applications (FCCFA) Aug-2022, CRC Press, ISBN: 9781003188230.

[21] Praveen, P, Mohammed Ali Shaik, T. Sampath Kumar, Choudhury T, “Smart Farming: Securing Farmers Using Block Chain Technology and IOT”, Aug-2021, EAI/Springer Innovations in Communication and Computing, ISBN: 978-3-030-65690-4

[22] M. A. Shaik, "Protecting Agents from Malicious Hosts using Trusted Platform Modules (TPM)," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 559-564, doi: 10.1109/ICICCT.2018.8473278.

[23] Amani Mobarak & AlMadahkah 2016, ‘Big Data In computer Cyber Security Systems’, IJCSNS International Journal of Computer Science and Network Security, vol. 16, no. 4, pp. 56-65.

[24] lan Gray & Mads Haahr 2004, ‘Personalised, Collaborative Spam Filtering’, in proceedings of the First Conference on Email and Anti-Spam (CEAS), Mountain View, CA, USA, July-August, under grant no. CFTD/03/219.

[25] Aakash Atul Alurkar; Sourabh Bharat Ranade; Shreeya Vijay Joshi; Siddhesh Sanjay Ranade, Piyush A Sonewar, Parikshit N Mahalle & Arvind V Deshpande 2017, ‘A proposed data science approach for email spam classification using machine learning techniques’, Internet of Things Business Models, Users, and Networks, pp. 1-5.