Efficient Modelling and Analysis on the Propagation Dynamics of Email Malware Filtering for Sustainable IT Development

[1]D.R.Thirupurasundari, [2]P.GururamaSenthilvel, [3]D.Sudha, [4]A. Hemlathadhevi, [5]Rajesh Kumar.K

[1,3,4]Assistant Professor, Department of Computer Science and Engineering, Meenakshi College of Engineering, Tami Nadu, India.
E-mail: tpsdr@yahoo.com,dsudha.1978@gmail.com,hemlathadhevi@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, Meenakshi College of Engineering, TamilNadu, India. E-mail: gurupandian.cse@gmail.com
[5]Assistant Professor, Department of Information Technology, Meenakshi College of Engineering, Tami Nadu, India.
E-mail: rajeshkumar07031990@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

ABSTRACT

Email system is one of the most effective and commonly used sources of Communication. Unfortunately, the Email system is getting threatened by spam emails. Email spam, also known as junk email or unsolicited bulk email (UBE), is a category of electronic spam which includes almost identical messages sent by email to multiple recipients. In existing system highly parallel encoding technique was used to detect the spam campaigns. Privacy preserving collaborative spam detection is used at the Receiver side. In the proposed system the modern spam filtering techniques are deployed at Sender Side. To increase the internet bandwidth and storage of the mail server to deploy WorldNet and Spambot. In this paper to use bro Intrusion detection to analyze the network traffic. It counts the number and regularity of the same email sent from a known IP address to various routes, and identifies disruptions on the network. Because of the large number of emails reported in the SMTP sessions, we efficiently monitor and process them in the Bloom filters. The process of spam filtering classified the email data into spam and harm mails.

Keywords: Spam Filtering, WordNet, Bloom filter, Sustainable IT Development, Propagation Dynamics.

1. INTRODUCTION

A network is an interconnected system composed of a number of human contexts (such as the organs of individual people). Most online social service providers, such as electronic mail and sending messages, are web-based and provide means for people to access over the web. Web-based social networking platforms support individuals who support activities and interests to be connected across political, economic, and geographic boundaries. Several other social media sites have additional features including being able to build communities that share similar interests. Internet email is among the most common mediums of expression in our corporate and personal interests. Spam messages are now a constant issue in email systems these days. Spam emails tamper with suppliers and end-users of both email services. Spam email involves almost identical messages sent by email to different receivers. Blank spam can also happen when a spammer is missing or otherwise failing to attach the message when the spam run is started. A spam filter should be customizable and simple to use. A more precise filter produces fewer false positives and less false negatives. False positives are valid emails which are wrongly perceived to be spam mails. False negatives are anonymous spam addresses. There are two major types of attacks on spam filters: poison attacks and attacks on the impersonation. A lot of legal words are applied to phishing emails in a poison attack, thereby reducing the likelihood of being identified as spam. A spam me imitates the identifications of ordinary users by establishing their IDs or jeopardizing their machines in an impersonation attack.

Spam filtering analysis can be categorized primarily into two subgroups: content based and identity based strategies. Emails are processed and scored in the content-based category based on keywords and patterns typical of spam. The simplest techniques to identity-based spam filtering are blacklist and white list, that also verify for spam detection by email addresses. White lists and blacklists also keep a register of addresses of individuals whose emails the spam filter does not and does reject, respectively. Spam filters based on email senders' identities. Most of the latest solutions for phishing detection are implemented on the receiver side. These protocols are good for detecting end-user spam words, butspam my links are consuming Internet bandwidth and memory resources. To identify spam bots, just use a monitoring system to control SMTP sessions and record the amount and similarity of receiver email addresses for each individual internal host 's outgoing mail emails as the spam bots detection features. Because of the large number of email related deals affecting the SMTP times, shop the information and financial within the Bloom filters along with dealing with these men. One main problem in today's Online Social Networks (OSNs) is to provide users the ability to monitor messages shared on their own private space to prevent posting inappropriate information.

This is accomplished via an adaptable rule-based system which enables users to make the filtering requirements that must be implemented to their wall. The automated program, called Filtered Wall (FW), is capable of removing abusive emails from user walls on the Online Social Network (OSN).

2. RELATED WORK

In 2019 Abdelrahman Al Mahmoudet al proposed a big data framework for collaborative spam detection as a spamdoop. Highly parallel learning algorithm used during spam campaign diagnosis. Collaborative detection of spam-conscious privacy is often used to discover viruses, worms, trojans and spams and then give messages to the recipient. It works favorably toward spam generation tools being generated and distributed overhead. It incorporates spamdoop systems, obfuscators, parallel classifier detector phenomenon[1].

In 2018 Sreekanth Madisetty, Maunendra sankar desarkar discussed Neural Network-based Ensemble encounters for spam filtering on Twitter. Often types are classified to identify spam at the tweet level, and suitable machine learning techniques are implemented in the literature. Recently, methods of deep learning have shown fruitful results on many tasks of processing the natural language. To this end, we are proposing a tweet level set approach to spam detection. Deep learning models are developed based on convolutional neural networks (CNNs).The ensemble utilizes Five CNNs and one feature-based model. Each CNN requires multiple words embedding (Glove, Word2vec) to build the machine[2].

In 2018 Peter Christen, Thilina Ranbaduge, Dinusha Vatsalan, and Rainer Schnell discussed about Precise and fast cryptanalysis for Bloom filter based privacy–preserving record linkage. A popular method used in PPRL is encoding sensitive values in Bloom filters (bit vectors), which has the potential to enable estimated matching using the q-grams character. Bloom filter encoding oriented PPRL has proved accurate and scalable tolarge databases. Bloom filters used during PPRL are resistant to threats of cryptanalysis which may re-identify a few of the sensitive values embedded in these Bloom filters. Although previous such types of attack were sluggish and involved knowledge of different encoding parameters, we suggest a new, efficient attack that utilizes how values of attributes are encoded into Bloom filters[3].

In 2017 Saeedreza shehnepoor, Mostafa Salehi, Reza farahbakhsh, Noel Crespib discussed. A Network-based Spam Detection System for Online Social Media Feedback. Classifying these hackers and the spam material is a hot research area and while a large number of studies have recently been carried out for this reason, the methodologies put forward to date still scarcely detect spam reviews and none of them demonstrate the value of each form of extracted feature. In this study, a novel framework called NetSpam use spam features to model evaluation datasets as heterogeneous communication technologies to map the process of spam detection into a classification issue in such networks. Using the value of spam functionality lets us get good outcomes from the Yelp and Amazon websites in terms of various metrics tested on real-world review datasets[4].

In 2015 Sanjeev das, Yangliu, Wei zhang, Mahintham Chandra mohan discussed Semantics-based Online Malware Detection toward more Efficient Real-time Protection against Malware. In this research we are proposing GuardOL, a hardware-enhanced architectural design for online malware detection. GuardOL is a hybrid Processor and FPGA solution. Our method is aimed at catching malware's malicious activity (i.e., high-level semantics). To this purpose, we first suggest the frequency-centered model for function development using patterns of established malware and samples in the device call. In FPGA we then create machine learning techniques (using multilayer perceptron) to use these algorithms to train classifier. The trained classifier is used for runtime to identify the unknown samples as malicious code or benign, with earlier detection[5].

3. METHODOLOGIES

Spam detection has been focus of considerable research[6]. Here we concentrate on collaborative approaches in which participants participate in the process of detection rather than results that show that the majority others of their spam detection. In addition, as the partnership provides new data sources, systems are involved in processing vast volumes of data in a limited time frame. Classic machine learning strategies such as Naïve Bayesian[7]. Linear discriminate analysis[8] and Support Vector Machines[9] confirmed the efficiency of this segment of spam detectors[10-15]. Such classificators have been extensively researched and contrasted in the research. Due to the emergence of open source data sets such as R, Bayesian studies are highly popular in the spam detection domain.

Effective implementation even so did not include any explicit parallelization control. Later, work was carried out on the implementation of machine learning techniques for spam detection purposes over Map Reduce. Cosdesis a spam detection framework which focuses on e-mails' HTML content / tags to conduct near duplicate similarity. Today, repositories which provide scalable scheme based of common machine learning techniques such as Apache Mahout are accessible. Those approaches, however, do not take into account issues of privacy. The paper introduces a method for carrying out broad scale spam analysis from various sources. Raw emails are gathered and analyzed using an implicit Map Reduce platform called OrientDB. The authors even so report that their distance calculation may take quite some time (upto several days). The report also does not comply with protecting email privacy.

Other preceding works and procedures claim to accomplish collaborative tracking of spam in a privacy-aware manner. The idea of Distributed Checksum Clear Housing (DCC) has been around for quite long time in particular. A DCC is a central network in which members exchange hashes from their emails. Then the application counts the number of times similar emails appear and marks suspicious ones as spam. The framework then counts the amount of times comparable emails appear and tags potentially malicious emails as spam. The DCC method based on distance preserving hashing methodologies that calculate inter-hash distances to check whether the emails generating hashes are similar. Some seminal experiment employed P2P mail server networks that exchange distance-preserving hashes to anonymously exchange spam information. In addition, a P2P design was used to enable e-mail users to review each other for comparable digests. More recently, a Large-scale Anti-spam Collaborative Privacy-Aware System (ALPACAS) has been suggested in and to inhibit inference attacks.

4. MALWARES AND EMAIL SPAM

There are many forms of email malware that can be sent to an infected device via email, Bots, trojans, viruses and a virus subset called a worm can send email messages to infected machines. As an attachment to an email message a worm can copy itself, sending itself to all your contacts. Most attachments to malware emails contain code or vulnerabilities that will allow your device to access more malware from the Internet. Email malware in recent years is also ransom ware that can erase or encrypt the files and backups even though they are saved in the cloud or on a computer. The answer is "no" for the vast majority of cases. You've likely heard of being hacked by email, so it's reasonable to worry that entering a hazardous email might get you tried to hack. They kept clicking on a malicious link in an email, or decided to open an attachment and decided to send it to them by email. Common types of spam mail,

- Spam of products and services: This is among the most prevalent offerings of a variety of products and services for routine use or for certain reasons irrespective of the language and geographical area probably change their products as per the season and circumstance with the goal of persuading users who are poorly prepared in a compelled and desperate manner, the spam traffic is extortionate and each time altering malicious strategy rises.
- Adult content spam: This spam is very prevalent because a proportion of normal users have to fall into this flashy trap for the material of products and services designed to improve adults' sexual life, it really should be mentioned that this is the second best-known spam mail which has been falling out of favor with the last 3 years but is still hazardous.

- Spam of health and medicine: This spam is recognized among several other things for its amount of offers in health supplements and skin care, often tugging more quickly than the previous ones and its incursion in the mail is very huge which is bad practice for the user or the corporation.

- Computer and Internet Spam: This spam is generally more hazardous than the prior ones since it has as its entity the commercial district providing hardware and software service providers of acquainted appearance to those who function in a company with deals for very under the framework of a computer service that makes it simple to realize the threat for users and businesses.

- Finance Spam: This is another treatment spam as its name suggests that it works on the banking sector, insurance plans and loans of low interest.

- Political Spam: It is well known for uploading files to inhabit the governorship or its management with public opinion polls about independent politicians etc.

- Virus in Spam mail: Usually, when a virus decides to invade the email via spam mail, there are many factors why the virus might attack each other because it can send enormously phishing messages packed from other viruses to various contacts seeking to broaden the infection or remove received emails in the inbox that are frequently used daily. But It's a reality that the device will alter doing any harm to the background once and seek to delete it from the first time instead of using the anti-virus.

- Ransom wares in mail Spam: There are several ways the Ransom wares can show itself off as an email spam now the modus operandi of cyber invaders with the release of Ransom wares is thru the distribution of crypto currency related advertising encouraging the consumer to download extensions or by reaching diffusion campaigns via WhatsApp, Instagram and even Facebook, which start running and playing until the device is locked once they are opened.

5. EXISTING SYSTEM

Spam mails in the Existing System filter on the receiver end. Typical sources include mail security software based on the cloud such as Symantec Message Labs and Google Postini, and also personal security product lines such as Kaspersky Internet Security and Avast Internet Security. Mail customers such as Microsoft Outlook and Mozilla Thunderbird, and also mail service providers, also endorse intrusion detection. The alternatives receive mail prior to actually filtering, so spamming actions still exist, and spam messages still end up wasting Internet bandwidth and the storage capacity of mail servers. Spam bots can obtain webmail interfaces, or provide spamming via secure SMTP. Because the messages are encrypted, the detection technique in this case cannot define the spamming bots. Spamdoop is a feature that enables multiple entities to work together to identify bulk malicious programs early. Our platform also satisfies the privacy requirements of participants.

The Obfuscator: Encodes the content of emails. The encoding enables concurrent spam filtering without violating the privacy of the original text. On the participant's side, the Spamdoop obfuscator is implemented to safeguard the confidentiality of the content of the emails. You can personalize the obfuscator to blend the company's objectives where supervisors can choose their favored cryptographic hashing feature. Also, it is easy to check the source code of the obfuscator to ensure that it maintains data integrity and confidentiality.

The Parallel Classifier: Uses the encoding characteristics to adjacent the process of forwarding digests correlating to similar messages to the same bucket. This element has main functions: it routes options linked digests to the same bucket and group data to be processed on the same bucket node. The properties of our encoding make this goal simple and inexpensive.

The Anomaly Detector: intercepts spam predicated on the huge growth rate of the buckets. Our method is premised on the theory that incidents and the rate at which spam messages arrive are very different from normal mail's. This is because spam is created by software tools that seek to make the most out of every sending as much of the messages as possible in a short possible time from the network spectrum.

To take advantage of this difference has adopted an anomaly detection technique based on histograms that has been used successfully for many other applications, including finding outlying instances in network traffic, or system calls in computers indicating compromised systems. The literature indicates quick and convenient histogram based anomaly detectors. To start with the construction of the e-mail occurrence density function in our histogram detection implementation; to calculate the number of messages that are present in the batch that many times for each number of occurrences.

6. PROPOSED SYSTEM

The spam filtering methods will be deployed on the sender side itself through the proposed model. By this the spam message cannot be sent to the receiver side. Spam email could also contain viruses as scripts or other connections to executable files. Typically some files come with an extension of.exe and encrypted file sent to the mails before submitting the package.The junk mail may also ignore the encrypted email using detection methods. This process will enhance memory storage and processing power. The encrypted format email, Word Net vocabulary and short message methodology are used in two approaches. The Bloom filters are being used to locate the junk mail, it has a benefit for reflecting sets on other data structures.This protects the detection and attempts to block spam messages through text based spam filters. And this is used to enhance Online Social service quality.

7. PROPOSED ARCHITECTURE

Spam mails on the recipient side are filtering on previous devices. Before filtering, the sender mail still exists so spamming practices, and spam messagesstill consume internet bandwidth and mail server storage space. Spam bots can access web mail interfaces, or provide spamming via secure SMTP. Since the messages are encrypted, this device cannot recognize the spamming bots by the detection method. Bayesian spam filters require a significant chance to adjust to a new spam based on customer feedback. For effective spam delivery, the spambot must distribute spam messages to a wide variety of innovative REAs. Because social networks have a lot of unwanted texts that used to be showcased on the wall in order to end up wasting memory space.

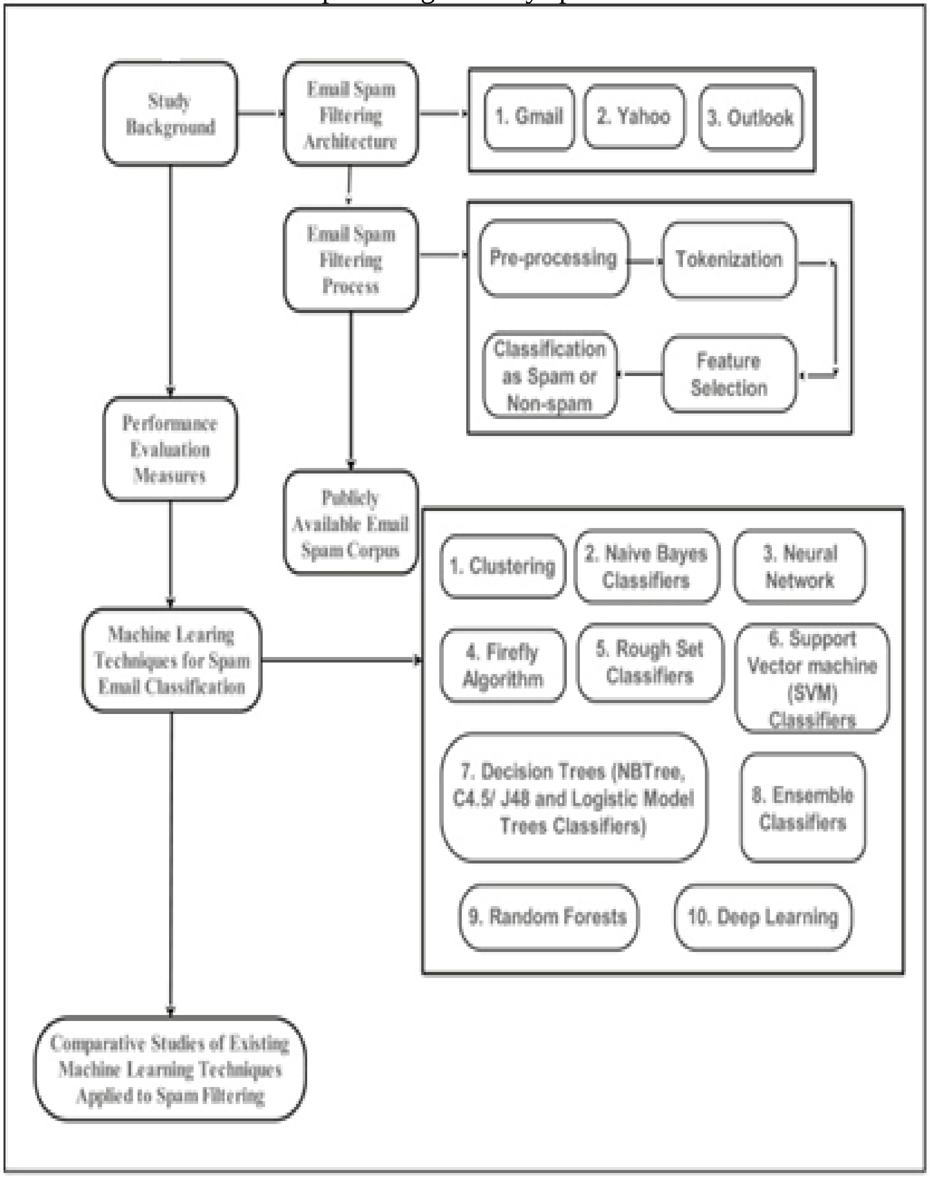

Study Background
Email Spam Filtering Architecture
1. Gmail
2. Yahoo
3. Outlook
Email Spam Filtering Process
Pre-processing
Tokenization
Feature Selection
Classification as Spam or Non-spam
Performance Evaluation Measures
Publicly Available Email Spam Corpus
Machine Learing Techniques for Spam Email Classification
1. Clustering
2. Naive Bayes Classifiers
3. Neural Network
4. Firefly Algorithm
5. Rough Set Classifiers
6. Support Vector machine (SVM) Classifiers
7. Decision Trees (NBTree, C4.5/ J48 and Logistic Model Trees Classifiers)
8. Ensemble Classifiers
9. Random Forests
10. Deep Learning
Comparative Studies of Existing Machine Learning Techniques Applied to Spam Filtering

Figure 1: Proposed Architecture- Detection of Encrypted Packet

Email Spam Filtering Process

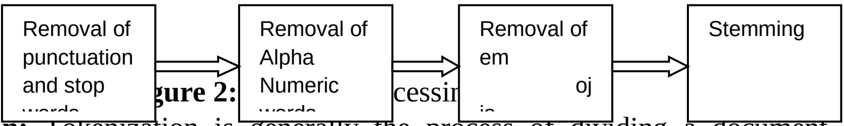
- **Pre Processing**: The pre-processing of text information is very important and important in filtering spam. The main purpose of text data preprocessing is to remove data that does not provide useful information about the document class. Also, we want to remove the redundant data.

The basic data pre-processing steps of spam detection are:
1) All the special characters are removed.
2) Stop words are removed.

3) Alpha Numeric words are removed using the ReGex..

3) Snowball stemmer Stemming Algorithm is applied to bring the word in their most basic form.

4) The word frequency of all the words are calculated.


Removal of punctuation and stop words
Removal of Alpha Numeric words
Removal of em oj is
Stemming

gure 2: cessi

- **Tokenization:** Tokenization is generally the process of dividing a document into parts, A significant part, or a token. These pieces or tokens is commonly referred to as a word or term. The creation of a type can be very essential or for separation of tokens. Tokens are mostly character items. A sequence of specific documents that are boxed together as a practical semantic module for the processing phase. This type is called the class of all tokens, including the same sequence of characters.
- **Feature Selection:** Entity selection is the process of selecting the subset of entities that are most relevant to your classification. We can increase the efficiency of the classifier by reducing the number of attributes in the feature space. We can dramatically reduce the time it takes to build and predict models. Optimal Feature Extraction (OFS): OFS is calculated based on information gain. Entropy (I) is used to calculate the uniformity of attributes and characterizes the purity (imperfection) of any collection of datasets. Information gain (G) is the expected reduction in entropy caused by partitioning the dataset.

- **Classification of Spam or not:** The Wordnet dictionary is a vocabulary database of semantic relationships between words. A word network used to detect unwanted words. For example: Rippling the surface - Emotional message (anger) I hate you - Emotional message (anger) - This message is blocked. Link Checker is a program that tests alphanumeric strings for unwanted video links (naughty videos). If a spam word or video link is detected, the sender will be alerted. Bloom filters are a space and memory efficient probabilistic data structure used to test whether an item is present in a set. The Bloom filter counts blocked emails. If an unwanted message or link is detected, the bloom filter will count it. Based on the count, you can monitor spambots based on the sender's network and the recipient's email address. Mail users send more spam messages to many recipients who can be identified by REA (recipient email address). Each device has a unique IP and MAC address that is used to access the Internet, so spambots can be identified by their network address. For example, checking the availability of usernames is a defined membership issue. Where the set is a list of all registered usernames. The price we pay for efficiency is that it is probabilistic in nature. This can lead to false positives. False positives can indicate that a particular username is already in use, but it is not actually in use. Repeated messages (advertisements) Network administrators and they cannot be sent to anyone. Based on REA and the sender's network, the network administrator monitors the number of blocks Network administrator if e-mails and count exceeds 10.Block spam bots. After blocking, spambots will no longer be able to access their spambots Account / system. After providing proper authentication to the Network Administrator, accounts blocked by the sender are released; we can access your account again.

8. RESULT AND DISCUSSION

In this module figure 1 shows a single individual can send one or more messages to another person, then join the two mail ids and subject bodies, add attachment and then click the send button. First, the bloom filter needs to check the subject and body texts, there is any wrong word or special symbols that are unrelated. If it's shown then block mail and increase count, Otherwise bloom filter will get the elements of the stream from the attachments and produce hash value for all streams. If there is no hash value for streams then obviously consider it to be undesirable files, block this mail as well. Users have many friends in their circle. If you want to notify any friend from this circle then chat / communicate via this network of e-mail communication.

Detection of Encrypted Packet: The mails and chat texts are also traveled in this module and then managed to reach the server. Files are encrypted for security purposes and then split as packets and network travel. After that, all the packets would be collected and re-assembled, then decrypted after it was stored on the server.

9. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, categorize the emails as spam or non-spam. With a large number of emails, unless people use the system it will be hard to handle all probable mails as our project deals with only a limited amount of corpus. The proposed model gives customers sensitivity and can very well adapt to changes. A primary aim of this paper is to provide a powerful and systematic technology to detect spam in the sender side. The Bro intrusion detection not only allows us to locate the spam sender but also allows us to block the spambot. The problem with spam email and anti-spam solution is cat and mouse game, as everyday spammers will come up with new email sending techniques. In the future, the study can be extended to block phishing mails and also broaden the denial of service attacks (DOS) to keep away.

REFERENCES:

[1] Abdelrahman Al Mahmoud, Ernesto Damiani, Hadi Otrok, Yousof Al-Hammadi, "Spamdoop: A Privacy-preserving Big Data platform for collaborative spam detection, IEEE Transactions on Big Data, Vol.5,no.3,pp. 293 - 304,2019.

[2] Sreekanth Madisetty , Maunendra sankar desarkar, "A Neural Network-Based Ensemble Approach for Spam Detection in Twitter", IEEE Transactions on Computational Social Systems, Vol.5, no.4,pp.973 - 984,2018.

[3] Peter Christen, Thilina Ranbaduge, Dinusha Vatsalan, and Rainer Schnell, "Precise and fast cryptanalysis for Bloom filter based privacy –preserving record linkage", IEEE Transactions on Knowledge and Data Engineering, Vol.31, no.11,pp.2164 - 2177, 2019.

[4] Saeedreza Shehnepoor, Mostafa Salehi, Reza farahbakhsh, Noel Crespi, "Net Spam: a Network-based Spam Detection Framework for Reviews in Online Social Media", International Journal of Innovative Technology and Exploring Engineering, Vol-8, no.-6, pp.748-752, 2019.

[5] Sanjeev das, Yang liu, Wei zhang, Mahintham Chandramohan, "Semantics-based Online Malware Detection: Towards Efficient Real-time Protection Against Malware", IEEE Transactions on Information Forensics and Security, Vol.11, no. 2, pp. 289 - 302, 2016.

[6] A. Khraisat, A. Alazab, M. Hobbs, J. Abawajy, and A. Azab, "Trends in Crime Toolkit Development", Network Security Technologies: Design and Applications: Design and Applications, 2014.

[7] D. Wang, D. Irani, and C. Pu., "A study on evolution of email spam over fifteen years", In 9th International Conference on Collaborative Computing: Networking, Application and Worksharing (Collaboratecom), IEEE, pp. 1–10, 2013.

[8] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy", Stanford InfoLab, Technical Report 2004-25, 2004.

[9] S. Y. Park, J.-T. Kim, and S.-G. Kang, "Analysis of applicability of traditional spam regulations to VOIP spam", Advanced Communication Technology, The 8th International Conference, Vol. 2, pp. 3 pp.–1217, 2006.

[10] P. Hayati, V. Potdar, A. Talevski, N. Firoozeh, S. Sarenche, and E. Yeganeh, "Definition of spam 2.0: New spamming boom", Proceedings of the 4th IEEE International Conference on Digital Ecosystems and Technologies (DEST), pp. 580–584,2010.

[11] D. Fallows, “Spam. how it is hurting email and degrading life on the internet”, the Pew Internet & American Life project, Washington, DC,USA, Technical Report, 2003.
[12] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, “Adversarial classification,” Proceedings of the tenth ACM SIGKDD International conference on Knowledge discovery and data mining, ser. KDD , pp. 99–108,2004.
[13] D. Chinavle, P. Kolari, T. Oates, and T. Finin, “Ensembles in adversarial classification for spam”, Proceedings of the 18th ACM conference on Information and knowledge management, ser. CIKM ’09. pp. 2015–2018,2009.
[14] B. Biggio, G. Fumera, and F. Roli, “Evade hard multiple classifier systems”, Applications of Supervised and Unsupervised Ensemble Methods, ser. Studies in Computational Intelligence, Vol. 245, pp. 15–38,2009.
[15] C. Pu and S. Webb, “Observed trends in spam construction techniques: A case study of spam evolution”, Proceedings of the Third Conference on Email and Anti-Spam (CEAS), 2006.