Malicious Url Prediction Using Machine Learning Techniques

Harsha Vardhan Sai Aalla
Computer Science and Engineering
Srm Institute of Science and Technology
Chennai, India
aa7673@srmist.edu.in

Nikhil Reddy Dumpala
Computer Science and Engineering
Srm Institute of Science and Technology
Chennai, India
dd7385@srmist.edu.in

Aswathy K. Cherian
Computer Science and Engineering
Srm Institute of Science and Technology
Chennai, India
aswathyc@srmist.edu.in

ABSTRACT

Phishing attacks have been a constant problem for years, despite diminution efforts from industries and the academic side. There were many attacks caused due to the insecure behavior. We accept that users fall into the trap of these websites because of a lack of education on it and not aware of these security threats and visible un abnormalities on the webpages they visit. We can also say that smart gadget users fall into this trap even more than computer users due to the screen size and performance. Most of the users fall into a hitch due to opening unnatural links and responding to unfamiliar recipients. Our project will help the users detect these kinds of phishing websites and reduce the risk of falling into this trap, and it can be useful for mobile phone and desktop users. Thus, we implemented a lightweight algorithm for detecting a spamming website without user activity. We used smart eye glasses to experimentally identify the mental efforts and vigilance shown by users while browsing a website and while playing a phishing game that we developed. We used fake login details as validation agents. We used some of the authorization agents and looked into the HTTP responses to determine the authorized webpages.

Keywords—Classification Techniques, Machine Learning, Spam Detection, Spam Filtering.

1.Introduction

Several users purchase products online and make payments through various websites. Multiple websites ask users to provide sensitive data such as username, password or credit card details, etc., often for malicious reasons. These types of websites are known as a phishing website. Phishing attacks have been increasing over the years. Simultaneously, several strategies have been introduced to reduce the attacks on users, but preventing attacks has been a significant challenge due to advanced techniques. Most mobile phone users are falling into this trap than desktop users due to computational powers and screen size. The attackers are launching these techniques through emails and instant messages. The attackers trap the users by creating a virtual environment and make them click on malicious links and downloading malware content into their smartphones and computers, in which the attacker is purporting to be a trusted contact. Cyberattacks are attacking all categories of people. The prime target of the attackers is social media websites, where users fall into the trap easily. However, attentiveness is as vital as cybersecurity knowledge. People need to know how the cyberattacks are taking place and beware of the facts.

Feature selection is one of the essential tasks which would be used when building machine learning models. Classification tasks is related with predicting a category of a data (discrete variables). An elevated level of understanding is required in a decision-making process; it can help project goals and interpret the results.

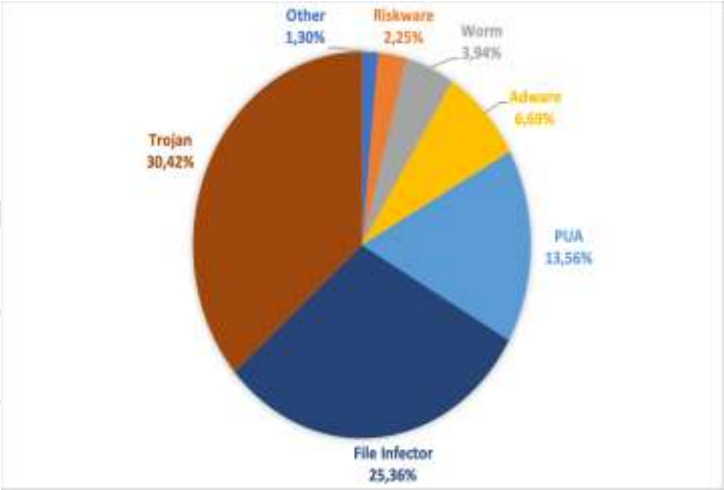

Fig 1: Different Types of Phishing Attacks

1.1 Machine Learning

Machine learning is a scorching topic for many vital reasons. It provides the ability to automatically obtain deep insights, recognize unknown patterns, and create high-performing predictive models without requiring explicit programming instructions. This high level of understanding is critical if ever involved in a decision-making process surrounding machine learning usage, how it can help achieve project and business goals, machine learning techniques to use potential pitfalls, and interpret the results.

1.2 Machine Learning Applications

Image Recognition: It is one of the most booming applications of machine learning. There are many cases where you can classify the object as a digital photo. For digital print, the measurements describe the outputs of each pixel in the photo. In the scenario of a black and white image, each pixel's intensity act as one measurement. So if a black and white image possesses N*N pixels, the total number of pixels and height is N^2. In the colored photo, each pixel is considered as three measurements of the intensities of 3 color components, i.e., RGB. So for N*N colored photo there are 3 N^2 measurements. For face detection– The categories are facing versus no face exist. There might be a particular category for each and every person in a individual database. For character recognition — We can divide a writing piece into smaller photos, containing a single character. These categories consist of the 26 letters of the English alphabet, the ten digits, and some special symbols.

Speech Recognition: Speech Recognition is the translation of words spoken into text. It is also known as automatic speech recognition, speech to text, or computer speech recognition. In speech recognition, a software application recognizes spoken words. The measurements in this Machine Learning application might be a set of numbers that represent the speech signal. We can segment the signal into portions that contain different words or phonemes. We can describe the speech signal by the energy or intensity in various time-frequency bands in each segment. Besides

the details of signal representation are outside the program's scope, we can represent the signal by a set of real values. Speech recognition, Machine Learning applications include voice user interfaces. Voice user interfaces are such as call routing, voice dialing, domestic appliance control. It can also use for simple data entry, speech-to-text processing, preparation of structured documents, and plane.

Extraction: Information Extraction (IE) is the critical application of machine learning. It derives structured information from unstructured data—for example, web pages, e-mails, reports, articles, and blogs. The relational database stores the output generated by the extraction of information.

The extraction process takes input as a set of procedure structured data and documents. This output is labeled as the table and an excel sheet in a relational database. Nowadays extraction is playing a vital role in the big data industry. As we know that a vast amount of data is getting generated, out of which most of the information is not structured. The first key challenge is handling the data, which is not structured. Now conversion of unstructured data to structured data based on the specific pattern so that the same can be stored in RDBMS. In addition to this, the data collection process is also getting change. Earlier, we used to collect data in batches like End-of-Day (EOD), but now the business wants the data as soon as it gets generated, i.e., in real time.

Prediction: Consider the scenario of a bank computing the probability of any loan applicants faulting the loan repayment. To calculate the fault's possibility, the system will need to classify the data available in particular groups. It is reported by a set of rules defined by the analysts. Once the classification is done, as per need, we can compute the probability. These probability computations can add across all sectors for varied purposes. The current prediction is one of the booming algorithms of machine learning. Let's take a shopkeeper scenario; earlier, we could get insights like sales report last year/month / 10-years / Holi / Diwali. This type of process is called historical reporting. But currently Business is more interested in finding out what will be my sales next year/month / Holi, etc. Business can take a crucial decision on time.

Classification: Classification is a method of dividing and placing each individual from the population under case study in many classes. Classification helps analysts to use object measurements to identify the type of class where the object resides. To provide an efficient rule, analysts use data. Data consists of many things with the correct classification.

In case of a scenario where a bank decides to disburse a loan, it assesses customers' ability to repay the loan. By considering customer's savings, financial history, earnings, and age, we can do it. This information is grabbed from the past data of the loan. Hence, Seeker uses it to create a relationship between related risks and customer attributes.

Machine Learning Tasks:

Clustering: Clustering is all about finding natural groupings of data and a label according to each group (clusters). The most common example includes product feature, customer segmentation, recognition for the product roadmap.

Regression: Regression tasks mainly deal with the prediction of numerical values. Some of the examples include the product price, estimation of housing price, stock price, etc. Some of the Machine Learning methods used for solving regression problems are kernel regression, regression tree, Gaussian process regression, Linear regression, Support vector regression.

Dimension Reduction: Dimension reduction reduces the number of random variables under consideration and can be divided into feature extraction and feature selection. Machine Learning methods used to reduce the dimensions are Manifold learning, Principal component analysis, Independent component analysis, Compressed sensing, and Gaussian graphical models.

Density Estimation: Density estimation problems are related to finding frequency or the likelihood of objects. In probability and statistics, density estimation is constructing an estimate function, based on observed data, of an unobservable underlying probability density function. Machine Learning methods used for solving density estimation tasks are Kernel density estimation, Density estimation tree, Mixture of Gaussians.

Multivariate Querying: Multivariate querying is about finding or querying similar objects. Machine Learning methods used for such problems are Nearest neighbors, Farthest neighbors, Range search.

Testing and Matching: Testing and matching tasks relate to comparing data sets. The methods used for such kinds of problems are Bi-partite cross-matching, Minimum spanning tree, and N-point correlation.

1.3 Machine Learning Classification:

Supervised Learning: These algorithms are trained using labeled data, in multiple scenarios, as an input where the outcome is already known. A piece of equipment, for instance, could have data points such as "F" and "R" where "F" represents "failed" and "R" means "runs."
A learning algorithm will receive some instructions from the input, along with the corresponding precise outcomes. The learning algorithm will then match the actual output with accurate development and show an error if there is any mismatch. Using different methods, such as classification, regression, gradient boosting, and prediction, supervised learning uses different patterns to predict a label's values on extra unlabeled data proactively. This method is commonly used in fields where historical data is used to predict events that are going to occur in the upcoming days. For instance, anticipate when a debit or credit card transaction is expected to be fraudulent or indicate which insurance customers can file their claims about issues**.**

Unsupervised Learning: This type of ML finds its application in fields where data does not consist of any historical labels. Here, the system will not provide the "right answer," The algorithm must predict the right answer. The main goal here is to analyze the data and identify a structure and pattern within the available data set. Transactional data acts as a well-defined source of data set for unsupervised learning mechanisms. For instance, this learning mechanism identifies customer segments with the same attributes and then lets them treat them similarly in business campaigns. Besides, it can also predict features that differentiate one customer segment from other customer segments. Either way, it is about predicting a similar structure in the available data set. Besides, these algorithms can also predict outliers in the available data sets.

Semi-supervised Learning: This type of learning is used and applied to similar scenarios, where supervised learning is applied. However, one must note that this learning technique uses both labeled and unlabeled data for the training process. Ideally, a small amount of labeled data, along with a large amount of unlabeled data, is used, as it takes a short time, money, and effort to acquire unlabeled data set. This machine learning technique is often used with methods such as classification, prediction, and regression. The companies that usually find it challenging to meet the high costs associated with labeled training methods will go for semi-supervised learning.

Reinforcement Learning: This learning technique is mainly used in robotics, navigation, and gaming. Actions that yield the best rewards are identified by algorithms that use trial and error methods. There are three significant components in reinforcement learning: the actions, the agent, and the environment. In this case, the activities are what an agent does, the agent is the decision-maker, and the atmosphere is anything that an agent interacts with activities. This kind of learning aims to select the actions that maximize the reward within a given time. By following an acceptable policy, the agent can achieve the goal faster and smoother.

Testing Process:

Testing: We all make errors, and some of them left unchecked; some of these mistakes can lead to bugs or failures that can be very expensive to recover. Testing our code helps to catch these errors or avoid getting them into production in the first place. Testing, therefore, is essential in the software development field. Tests help to identify bugs, ensure the product quality, and to verify the software.

Unit Testing: Unit testing is defined as the design of tests that validate that the internal program logic is adequately operated and that program input produces efficient outputs. All branches of choice and internal code flow should be authorized. It validates separate software units of the request .it is done after the close of an individual team before integration. This is structural testing that depends on the data of its structure. Unit tests achieve necessary tests at the factor level and test a specific commercial process and system formation. Unit tests ensure that every single path of business process completes accurately to the documented provisions and contains clearly defined inputs and probable results.

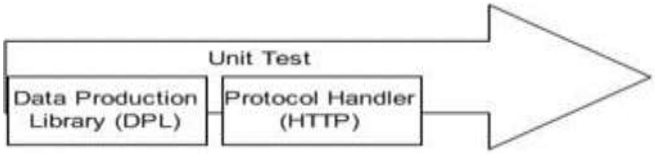

Fig 3: Unit Testing

Integration Testing: Integration tests are performed to test integrated software components to regulate if they run as one program. Testing is occasion driven and is more concerned with the immediate result of screens or fields. Integration tests validate that although the workings were individually approved, positive unit testing showed that the grouping of components is correct and dependable. Integration testing is specifically aimed at revealing the problems that arise from the collection of components.

White Box Testing: White Box Testing is a challenge in which the software tester has information about the software's inner workings, construction, and language, or at least its drive. It is used to validate areas that cannot be stretched from a black-box level.

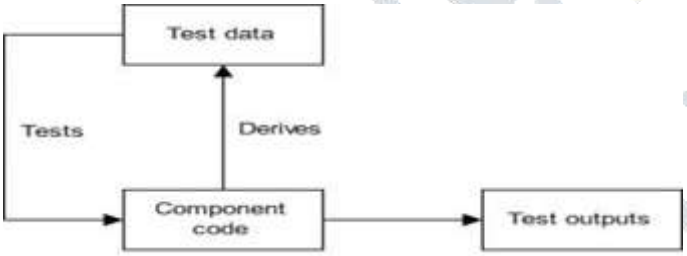

Fig 4: White Box Testing

System Testing: System testing confirms that the entire combined system software meets supplies. It tests a configuration to ensure predictable and known outcomes. An example of system testing is the oriented arrangement system mixing test. System testing is based on procedure similes and flows, emphasizing pre-driven process links and additional points.

Black Box Testing: Black Box Testing is a technique of testing the software short of any knowledge of the buildings, inner mechanisms, or language of the module actuality tested. As with most other kinds of tests, black box tests must be printed from a final source document, such as requirement or necessities file, such as specification or requirement file. It is a testing in which the software below test is treated as a black box, you cannot see into it. The test provides inputs and responds to outputs without seeing how the software works.

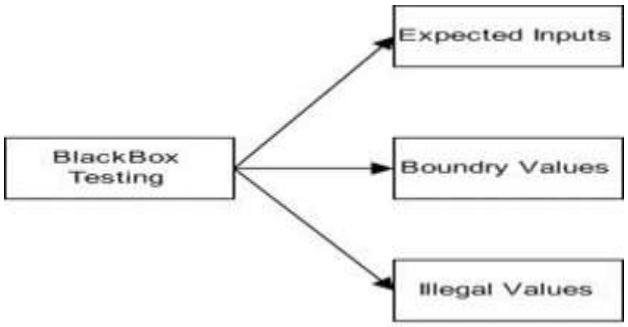

Fig 5: Black Box Testing

2.Literature Survey:

Many researchers have used various methods over the years for the detection of spamming websites. The following paragraphs give a quick insight into the works completed up to date.

(1) Phishing is a cyberattack that strengthens the social engineering approaches and other sophisticated related techniques to get personal information from website users. According to a re- cent survey, the annual growth rate of the number of unique phishing websites detected by the anti-phishing working group for the past six years is 36.29% and for the past two years is 97.36%. In the wake of this increase, alleviating malicious attacks has received a growing interest from the cybersecurity community. Extensive research and development have been detected based on their network, unique content, URL characteristics. The existing approach- es differ significantly in terms of data analysis methods, intuitions as well as evaluation methodologies.

(2) The technique used for robust and most efficient detection of phishing via page component similarity. Nowadays, social network platforms have become the most famous platforms for users to interact with different people. Given the sensitive information available on a social media platform has become one of the most important research issues. As a traditional information stealing technique, malicious attacks still work in their way to cause a lot of privacy violation issues. An attacker sets up spam web pages in a web-based vicious attack, pretending to be an original social media portal to make users input their sensitive information, such as social security numbers, debit card numbers, passwords.

(3) Most of the existing phishing identification techniques are not effective against domain name system-poisoning-based phishing attacks. Proposed is an effective method for identifying such attacks: websites' network performance features are used for classification. The performance of the four classification algorithms is explored to demonstrate the approach. Linear discriminant analysis, support vector machine, K-nearest neighbor, and naive bayesian. More than ten thousand real-world items of routing information are observed for one week.

(4) An fMRI study of malware and phishing warnings. The security of computer systems often depends upon the decisions and actions of end-users. In this survey paper, we set out to investigate user' susceptibility to cybersecurity attacks by focusing on the most fundamental component governing user behavior the human brain. We introduce a novel called neuro- science-based study methodology to inform the design of user centered systems related to cybercrime. In particular, we focused on a functional magnetic resonance imaging study determining users' underlying neural activity concerning two crucial activities security tasks: (1) heeding security warnings and (2) distinguishing between a phishing and a legitimate website and security performance.

(5) Phishing is the art of echoing a creditable firm website intending to grab user' s private information such as passwords, usernames, social security numbers. Phishing websites con- sist of various cues within its content-parts and security indicators related to browser-based provided along with the website. Several advanced solutions have been proposed to tackle

phishing. Nevertheless, there is no single bullet that can radically solve this threat.

(6) This survey paper shows a focused survey of Machine Learning (ML) and Artificial Intelligence (AI) methods for intelligent email spam detection, which we believe can help build appropriate countermeasures. Based on the relevance of a smart approach, papers representing important methods were read, identified, summarized.

(7) The spam review detection problem has gained much attention from communities and researchers, but still, there is a need to conduct experiments on large-scale, real-world review datasets. This can help to examine the impact of widespread opinion spam in online reviews. Experimental evaluations are performed on a real-world Amazon review dataset, which examines 15.4 million reviewers and 26.7 million reviews.

(8) This literature survey addresses the most challenging throughput issue and offers a constant time complexity rule-based spam detection algorithm. The algorithm has a static processing speed, which is independent of vocabulary size. A new emerging data structure, namely, Hash Forest, and a rule encoding process, are developed to make constant time complexity possible.

(9) Feature selection is an essential task in keyword content classification for being among the most popular and vital feature reduction methods. The cross-validation was used for the validation and training dataset, and seven datasets were employed in testing the spam classification proposed. The results demonstrate that the meta-heuristic, namely water cycle feature selection (WCFS), was engaged and three ways of hybridization with Simulated Annealing as a feature selection used.

(10) Twitter spam detection refers to a tricky task for the involvement of a range of characteristics, and non-spam and spam have caused the unbalanced distribution of data on Twitter. Twitter spam characteristics were analyzed as the content, user attribute, activity, and relationship in this study to solve the problems. A novel spam detection algorithm is designed based on a regularized learning machine, called the Improved Incremental—Fuzzy-kernel-regularized Learning Machine (I2FELM), useful to detect Twitter spam correctly.

3.PROPOSED WORK:

3.1 Data Process Diagram:

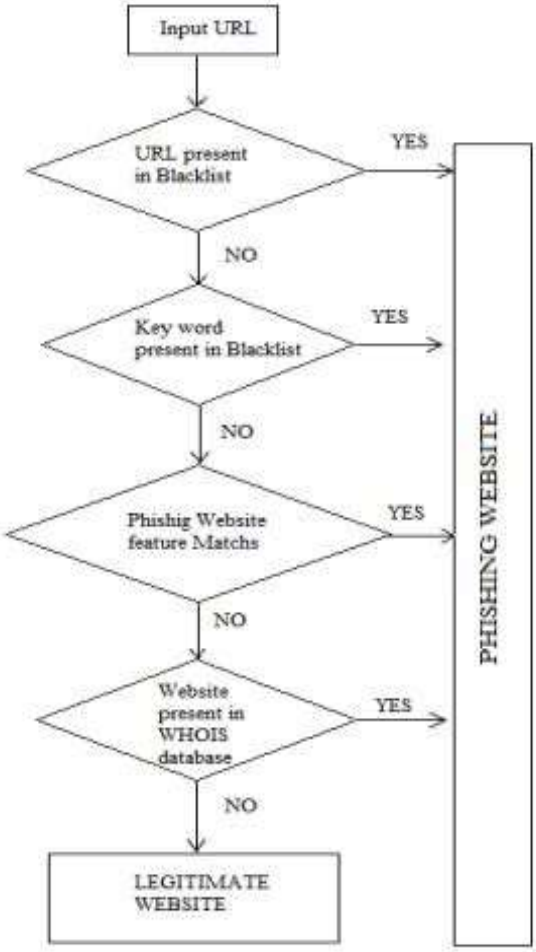

Fig 2: Data Process

1.Data Preparation and Pre-Processing: Data Preparation is one essential and crucial part of the machine learning model. Data Preparation involves data collection, selection, and pre-processing

1.1 Data Collection: Information based on some targeted variables to analyze and produce some valuable outcome. However, some of the data may be noisy, may contain inaccurate values, incomplete values or incorrect values incomplete values or false values. The foremost step in data collection is gathering malicious URLs from the WHOIS database, Phish Tank, and other sites. Our datasets are not the same because it is collected at different times. These platforms help developers to produce malicious reports and combine anti-phishing data into developed applications.

1.2 Data Selection: Data collection consists of attributes required for the project and it includes some functions or methods that allow us to select useful data, Selecting the required data helps in reducing cost of the model. The second foremost step is selecting data. The selected data should be feasible and useful data should be extracted from the datasets. This selection helps to contribute to the desired output, which is required. The selected data should be free from misleading data and less noisy.

1.3 Data Pre-Processing: The aim of pre-processing is to convert raw data into processed data that fits machine learning. Cleaning data helps to precise and accurate result when applied on machine learning. Data pre-processing includes cleaning of data, data normalization,

transformation, and feature extraction. The final product of data pre-processing is the last training set. The pre-processing of data will generally have an impact on the output and extracted data set. The pre-processing step is necessary to sort out several types of problems: noisy data, missing data values, redundancy data, etc. In all cases, missing data values should be pre-processed to process the Machine Learning algorithm's real data.

2.Data Splitting: Data Set needs to be spitted into three subsets which includes training, test and validation. The, more training of data gives better results which leads to better model performance and capability. Cross-Validation is a vital part of data splitting when the training of data. Some portion of data can be used to develop a model and improve the performance of the model.

3.Modelling: It feeds the algorithm with trained data set, and target output is obtained. Model tuning and cross-validation are done which improves prediction, which helps in improving model performance. Modeling of data is trained to recognize patterns of different types; a data set can be prepared by providing an algorithm and derive conclusions and learn from those data sets. Modeling is a process of training a machine-learning algorithm to predict values from the data set.

4.Model Refinement: Performing model-specific optimization and run error analysis, which helps in the development of an advanced model**.**

5.Testing and Evaluation: Evaluate the model based on the distribution in data sets and note the difference between trained and tested data set. The model evaluation aims to estimate the accuracy of the predicted data. Test data is used to test model performance. Accuracy is a standard evaluation metric for classification problems. It is the fraction of the number of accurate predictions to total predictions made.

Proposed Work Process: After initial analysis, the whole activity's computerization is being suggested by considering the anomalies in the existing system. The major drawback in existing systems is they are not working under real time environment. Our proposed work will help the people to check website is phishing website or not in real time environment. The authentication in mobile devices is more challenging than in desktop computers. The important automation drivers that are available on desktops are commonly not found on mobile devices. In this work, we automated the authentication procedure of users through cylindroid with java script object notation. To make the transactions secure, we have added some technical features to our system that will be used in eCommerce websites, and users will buy the products without any hesitation. Some persons want to know whether their devices are breached or not, but there are no proper checking resources. Our proposed system will provide a feature to check their devices and offer them security features to be aware of malicious sites. In the present scenario, due to an increase in technical advancements, spammers came with a new idea virtual website which looks like the original website. Still, the only change is the java script code running in the background. It isn't easy to find these types of websites. So, in our proposed system, as it is working under real time environment these kinds of websites are easily detected.

4.CONCLUSION:

The authentication automation in mobile phones is more challenging than in computers. The necessary automation drivers that are available on desktops are commonly not found on mobile phones. In this work, we automated the user authentication process through Selendroid with JavaScript Object Notation (JSON). However, Selendroid is only compatible with Android devices and does not work on other types of devices. We proposed automatic login with fake details to be deployed on browsers for phishing detection. Phishing sites tend to grant permission even with the incorrect login details, which requires less meager computational resources. It would not require user efforts, making it useable and sustainable by non-aware users and cybersecurity aware.

5.REFERENCES:

(1) Zuochao,Dou,Abdallah Khreishah (2017), Systematization of Knowledge (SoK) : A Systematic Review of Software-Based Web Phishing Detection

(2) Wenqian Tian,Zhenkai Liang (2017), Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity**.**

(3) Fadi Thabtah,Lee McCluskey (2014),Intelligent rule-based phishing websites classification.

(4) Jun Ho Huh,H.Kim (2011),Detection DNS-poisoning-based phishing attack from their network performance characteristics.

(5) Nitesh Saxena,Jose Omar Maximo (2016), Neural Markers of Cybersecurity: An fMRI Study of Phishing and Malware Warning.

(6) Asif Karim, Sami Azam, Bharanidharan Shanmugam, Krishnan Kannoorpatti, Mamoun Alazab (2019) , A Comprehensive Survey for Intelligent Spam Email Detection.

(7) NAVEED HUSSAIN, HAMID TURAB MIRZA ,IBRAR HUSSAIN, FAIZA IQBAL AND IMRAN MEMON (2020), Spam Review Detection Using the Linguistic and Spammer Behavioral Methods.

(8) TIAN XIA (2020), A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems.

(9) GHADA AL-RAWASHDEH, RABIEI MAMAT, AND NOOR HAFHIZAH BINTI ABD RAHIM (2019) Hybrid Water Cycle Optimization Algorithm with Simulated Annealing for Spam E-mail Detection.

(10) ZHIJIE ZHANG, RUI HOU1, AND JIN YANG2 (2020) Detection of Social Network Spam Based on Improved Extreme Learning Machine