isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 3 ,Apr,2024 ,ISSN_NO: 2321-3337

DYNAMIC IDENTIFICATION OF SPAM WORDS, PHISHING URLS, MISLEADING PRODUCTS AND AGENCIES.

Bhavani.R[1], Dhanalakshmi.K[2], Varshini.M[3], Malathi.S[4*]

[123]UG Scholar –Dept.CSE, GRT Institute of Engineering and Technology, Tiruthani , India.

[4*] Assistant Professor-Dept.CSE, GRT Institute of Engineering and Technology, Tiruthani , India.

bhavanirb901@gmail.com,dhanalakshmi141102@gmail.com,varsh233003@gmail.com

***Corresponding Author: s.malathi83@gmail.com**

Abstract

In general, Spam has become the platform of choice use by cyber-criminals to spread malicious payloads such as viruses and trojans. Collaborative spam detection techniques can deal with large scale e-mail data contributed by multiple sources; however, they have the well-known problem of requiring disclosure of e-mail content. In our Project, we are designing Social media like web page to filter the Spam messages that are getting posted. This will filter the irrelevant and spam posts so that wrong information can be avoided. Phishing URLs are also avoided in order to remove the duplicate websites. Wrong information can be removed based on the public votes. We are using Big Data and Block chain Technology for further Data analysis and data security process. Hadoop distributed File System (HDFS), is used for Big Data analysis. In Block chain, 4 different Algorithms are used namely, Asymmetric Key Algorithm, Digital Signature Algorithm, Secured Hash 256 (SHA256) Algorithm & Merkle hash Tree Algorithm.

Keywords: *spam detection, Phishing links, Big data, Hadoop distributed file System, Block chain.*

1. Introduction

Social platforms in which individuals can easily express themselves and share information have had a very important place in our lives. Social networks such as Twitter take a significant part in social communication among the people. The word spam was originally used to describe unsolicited e-mails sent in bulk. It is hard to define the term spam more accurately. Some argue spam is about the lack of consent on the part of the recipient, while others believe it is about unsolicited e-mail quantity or scale. Other definitions also stressed the commercial nature of spam for example, the US SPAM act of 2003 established stringent requirements for sending commercial e-mails. Later, spam got closely associated with cyber-crime. Spam e-mails often try to ensure the recipient to click on a fake or infected URL that links to a malicious Website (phishing) or downloads a malicious attachment containing a zero-day exploit (spear-phishing). Phishing is a type of cyber-attack that uses deceptive websites, or text messages to trick users into providing sensitive information. Web Phishing is a form of cybercrime where criminals attempt to steal sensitive information such as login credentials, credit card details and personal data by discussing themselves as a legitimate entity through a fake website or e-mail.

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. The challenges include analysis, capture, duration, search, sharing, storage, transfer, visualization, and privacy violations.

Block chain records all the transactions which are generated in a peer-to-peer network, and its is actually a decentralized ledger system. In the system, all the blocks include the hash of the previous block, in this way, they are linked together by the hash, and a block chain is

formed. According to the decreasing order of decentralization, the block chain consists of three categories: public block chain, consortium block chain, and private block chain.

2. Related Work

Bulk emails or spoof emails sent to a specific individual or entity are known as spam. Several approaches and machine learning techniques can be used to identify these spam emails. Actual private data loss may result from the spam emails. There are new methods for determining if an email is spam or ham. Spam is unsolicited mail, yet ham is legitimate mail. Researchers of today have included certain text messaging functionalities. These are employed to categorize them as spam or ham. This study evaluates the accuracy of several classification techniques by comparing them with data collected from various sources. Our NLP algorithm separates and categorizes emails as rubbish or spam. The Extreme Learning Machine (ELM) is an illustration of a learning-based model [1].

End consumers are put in danger of financial loss as well as health and safety when counterfeit or duplicate goods are manufactured and marketed. Through revenue loss, product defamation, downtime, replacement costs, and other negative effects, it also negatively impacts the economic growth of original manufacturers and enterprises, compelling many to spend money on legal defense. Counterfeits, a company partner's trust may be compromised; sales may be stolen, etc. A blockchain-based system is used to identify original products and also detects duplicate products to assure the identification of original goods in order to combat and stop these critical repercussions of counterfeiting. In this project, QR (Quick Response) codes and barcodes offer a way to reduce the practice of counterfeiting, given the rapidly developing trends in wireless technology [2].

Phishing attacks pose a serious risk to an individual's or an organization's security. Phishing URLs are created expressly to trick people into thinking they are authentic websites, which allows them to steal private data like credit card numbers, usernames, and passwords. Researchers have created a number of methods for spotting phishing attempts in order to combat this danger. URLs, the majority of these methods have significant false positive rates and poor detection rates. In this research, we provide a new method for phishing URL identification based on Support Vector Machine (SVM) techniques and machine learning [3].

Every year, unsolicited emails—such as spam and phishing emails—cost people and companies millions of dollars. Although many models and methods for automatically identifying spam emails have been created, none of them have demonstrated 100% predictive accuracy. Both machine learning and deep learning algorithms outperformed all other models that were put out. Natural language processing (NLP) improved the accuracy of the models. This study presents the efficacy of word embedding in the classification of spam emails. BERT (Bidirectional Encoder Representations from Transformers), a pre-trained transformer model, is optimized to perform the task of distinguishing spam emails from non-spam (HAM). BERT employs attention layers to put the text's context into context. The outcomes are contrasted with a basic deep neural network (DNN) model that has a bidirectional long short [4].

Recently, there has been an increasing trend on the Internet, especially with Online Social Media (OSM) platforms, such as Facebook, Twitter, and others, which are turning into massive information repositories. Because it is submitted by users on these websites, the content is large, disorganized, erratic, and dynamic by design. In addition to legitimate users, spammers and users who want to disseminate harmful or irrelevant content are often reported to be rather active. We concentrate on Twitter spamming in our work. Users of Twitter, referred to as reporters, are

usually the ones that report spam activity, and those who engage in spamming activities are referred to as reporters. We gathered information about alleged spammers, or reporters [5].

3.Objective

Our project aims to deliver a comprehensive solution that not only enhances users' social media experience but also effectively combats fraudulent activities, ultimately fostering a safer and more secure online environment for all users. Harnessing big data analytics to analyze patterns and trends indicative of spam, phishing URLs, and counterfeit products on social media platforms. Developing Block Chain algorithms to proactively detect and intercept malicious content, including spam, phishing URLs, and counterfeit products, in real-time.

4. Proposed System

In the proposed system, We will train the system with set of spam keywords and Phishing URLs in the backend server. In the front end we will designing Social media like web page to filter the Spam messages that are getting posted. We also integrate Big Data – Hadoop Distributed File system (HDFS) to filter the Spam keywords, phishing URLs and irrelevant contents from our Prototype social media. Phishing URLs are also avoided in order to remove the duplicate websites. Wrong information can be removed based on the public votes. Spam posts are removed so that irrelevant posts are removed from our social media. Phishing url's are identified in order to identify the duplicate and malicious url and to avoid internet frauds and also we detect the fake products and agencies based on the feedback and reviews by the other customers.
The problem of early detection of Collaborative spam detection can deal with large scale e-mail data contributed by the multiple sources.

5. Architecture Diagram

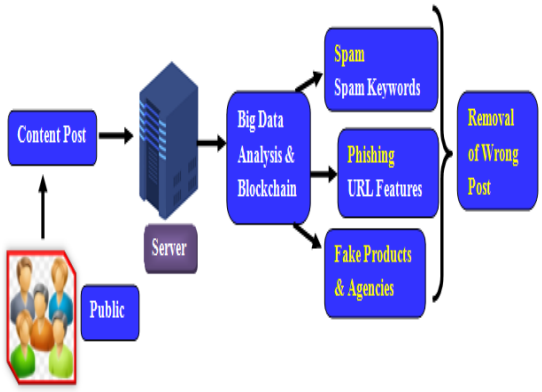

Fig.5.1.Architecture Diagram

The architecture diagram of the social media platform as shown in this consists of three main phases posting the content, sending it to the server, validating and removing of contents

6. Algorithm

Big data is used as the main analytical tool in this project. Big data- Data sets with sizes beyond the ability of commonly used software tools. The most fundamental challenge for Big data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In Fig.6.1 we described data volume, data velocity, data variety, data veracity.

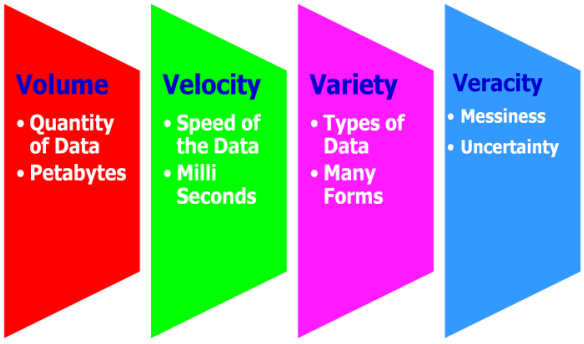

Fig.6.1. 4V's of Big Data

isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 3 ,Apr,2024 ,ISSN_NO: 2321-3337

4V's of Big Data:

Data Volume:

1.44 x increases from 2009 2020
2. From 0.8 zeta bytes to 35zb
3. Data volume is increasing exponentially

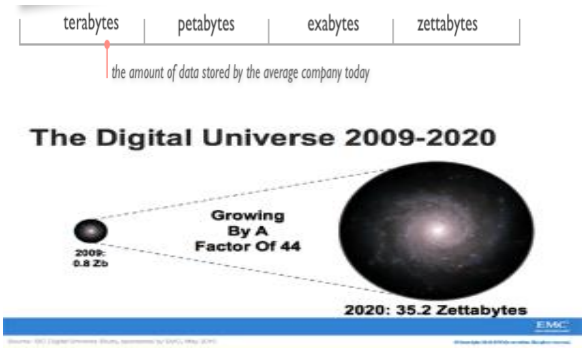

Fig.6.1.1. Data Volume

Data Varity:

1.Various formats, types, and structures
2.Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
3.Static data vs. streaming data
4. A single application can be generating collecting many types of data.

Data Velocity:

1.Data is begin generated fast and need to be processed fast.
2. Online Data Analytics.

Veracity:

Data will be mess and Uncertainty.

7. Implementation

7.1 User Interface Design

In this Module, user interface module is designed using web application using Java. The user has to be registering User details. Such as Name, Mobile number, E mail ID and other credentials. All the data provided by the user is stored in the centralized server.

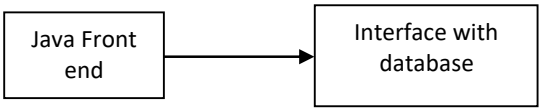

Fig. 7.1 User Interface Module

7.2. Server

In this module, all the Data are stored in the centralized server. All the user information as well as private credential where stored in the centralized server. Whenever it is needed all the data or all the information is used for verification / authentication. We use MySQL as the database part.

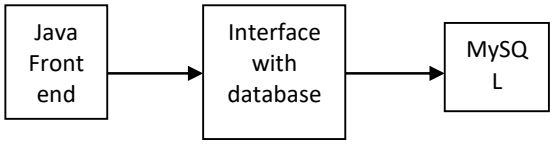

Fig.7.2. Server Module

7.3 Spam Keywords Training

In this module, Spam keywords are stored and kept for data analysis and comparison. The Spam words can be added whenever it is needed for fetching more accuracy. All the spam keywords are used for comparison component of analysis from the user input. This module helps to train the spam keywords so that our application detects and separates the spam contents from the posting made by the user.

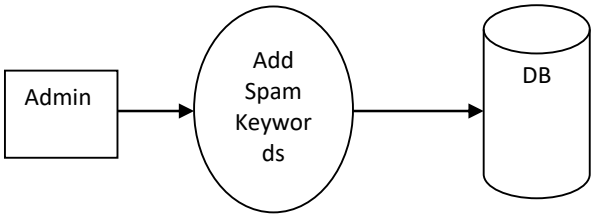

Fig.7.3. Spam keywords Training

Volume: 6, Issue: 3,Special Issue: 3 ,Apr,2024 ,ISSN_NO: 2321-3337

7.4 Spam Detection

In this module, after training of spam keywords the server automatically detects the spam words which are present in the content posted by the user. We deploy detection module through Naïve Bayes Classifier algorithm. Naive Bayes classifiers are a popular statistical technique of e-mail filtering. They typically use bag of words features to identify spam e-mail, an approach commonly used in text classification.

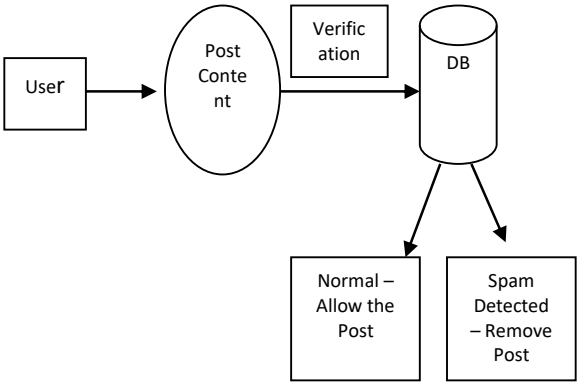

Fig.7.4. Spam Detection

7.5. Phishing Url Detection

In this module, Holistic approach is used for detection of phishing URLs. We implement this approach by comparing with the Dataset. Once we find the requested URL is present in the phishing set URLs then the requested URL is blocked by the server.

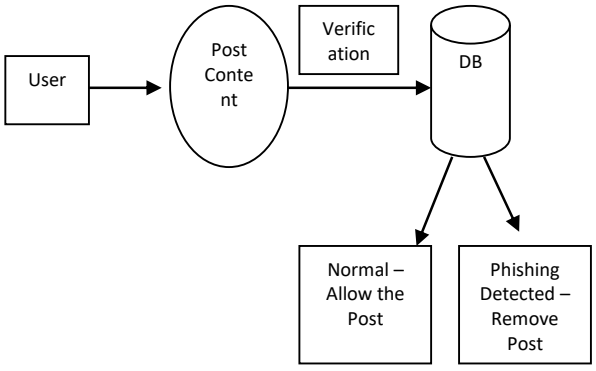

Fig.7.5. Phishing URL Detection

7.6. Fake Products & Agencies Detection

In this module, Fake products and fake Agencies are detected through the multiple feedbacks posted by the different users into the server. User's Feedbacks are analyzed through our application and fake products and agencies are detected and notified to the new users. Based on the feedbacks fake products and agencies can be omitted from the list, so that user may not be shown with those fake products and agencies.

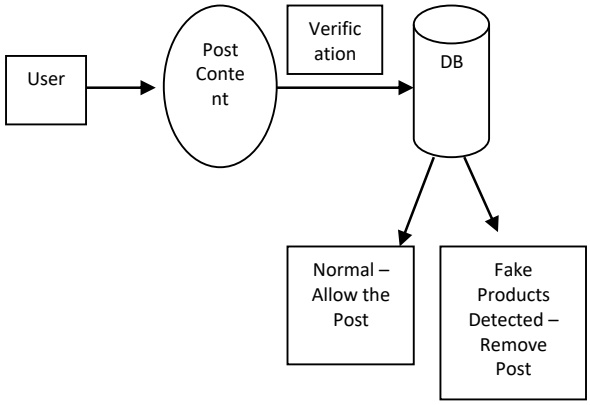

Fig.7.6. Fake Products Detection

8. Experimental Results

This result discusses about the detection and removal of the Spam words, Phishing URL and Misleading Products are detected and removed.

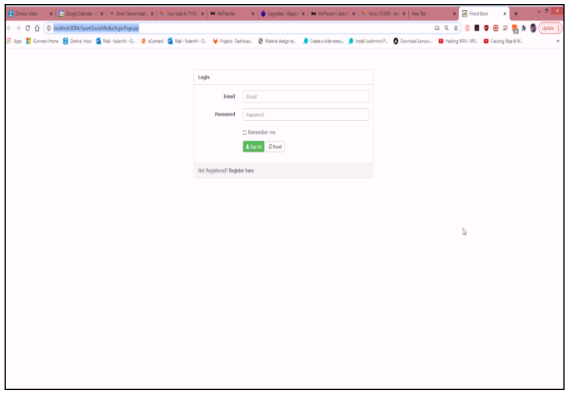

Fig.8.1. User Sign In Page

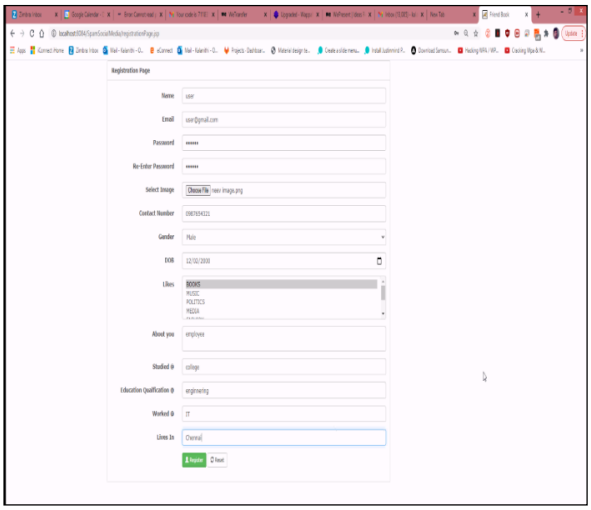

Fig.8.2. User Registration Page

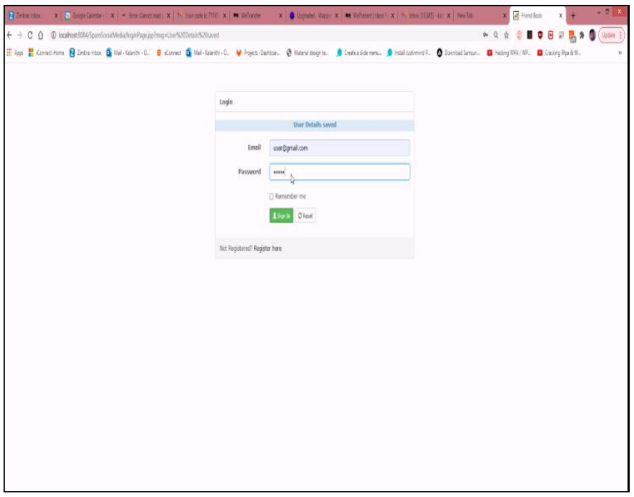

Fig.8.5. Shows the Login page

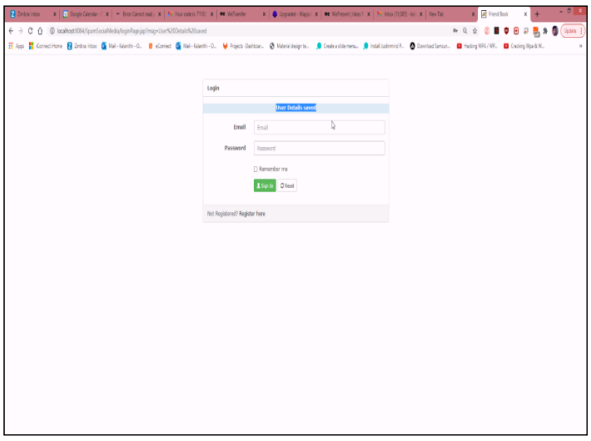

Fig.8.3.User Login Page

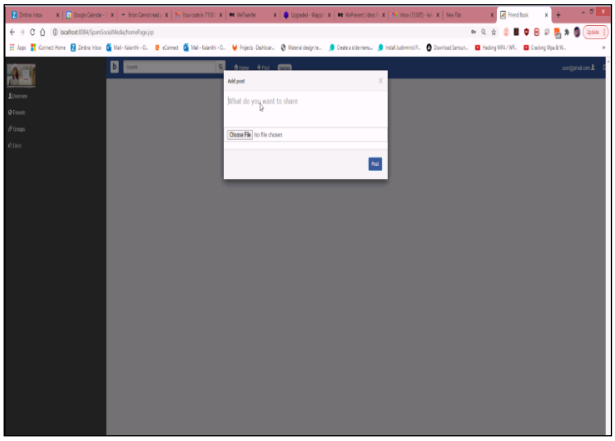

Fig.8.7. Shows the webpage

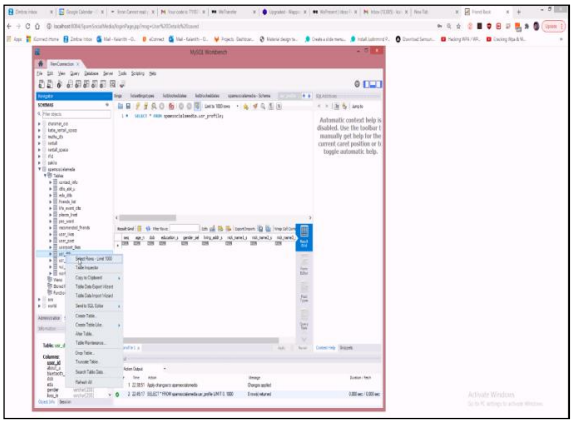

Fig.8.4. Data Structure Schema

9. Conclusion & Future work

Our project endeavors to combat spam and phishing threats on social media by implementing a robust detection and filtering system. We Have deployed social media like web page to detect spam words and the malicious URLs and also dealers so as to ensure proper and genuine product sales through social media we also identify the fake dealers and fake products through this system. We deploy android based application to cluster the messages based on keywords. Through the comparison of user posts with a trained dataset, we swiftly identify and remove any

content containing spam keywords or phishing URL's ,fake products and ensures the integrity and confidentiality of user data. Through our dedication to detection, prevention, and security, we strive to deliver a reliable and secure platform for users to engage and interact online. So we aim to create a safer and more trustworthy environment on our Prototype Social Media platform. In Future Multiple Phishing URLS and Spam Keywords are integrated into our Application and lots of training sets are added together and trained using Machine Learning Techniques.

Reference

[1]Abishek Sharma, Arjun N. Spam Detection Using Machine Learning Techniques Network Security, Vol:4, Pages 1 – 11, 2023.

[2]Nruthya Ganapathy B , Keerthan Kumar, Poojary Shreya Jaya, Rajath D Shetty, Dr. Shreekumar T, Fake product detection using blockchain technology ,Vol:10, pages 1–5, 2022.

[3]Mrs.Sarika Dhurgude, Varun Awargaonkar, Omkar Doiphode, Pratik More,Abhijeet Shilawant, phishing url detection using machine learning, Vol:11, pages 1–3, 2023.

[4]Isra'a AbdulNabi, Qussai Yaseen, Spam Email Detection Using Deep Learning Techniques, http://creativecommons.org/licenses/by-nc-nd/4.0/, pages 1–6, 2021.

[5]PoojaSinha, OshinMaini, Gunjan Malik and RishabhKaushal , Ecosystem of Spamming on Twitter: Analysis of Spam Reporters and Spam Reportees Vol.13, pages 1–7, 2022.

[6]G. Cormack. Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval, Vol. 1: Pages 335–455, 2007.

[7]A. Khraisat, A. Alazab, M. Hobbs, J. Abawajy, and A. Azab. "Trends in Crime Toolkit Development" in Network Security Technologies: Design and Applications: Design and Applications. IGI Global. Ch-2, pages 28-43, 2014.

[8] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The underground economy of spam: A botmaster's perspective of coordinating large-scale spam campaigns. USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), Vol. 11: Page 4, 2011.

[9]M.Crawford T. Khoshgoftaar J. Prusa. "surve of review spam detection using machine learning techniques". Journal Of Big Data, Vol. 2, pages 23, 2015.

[10] M. Sheikhalishahi, A. Saracino, M. Mejri, N. Tawbi, and F. Martinelli. Fast and effective clustering of spam emails based on structural similarity. In International Symposium on Foundations and Practice of Security, pages 195–211. Springer, 2015.

[11] J. Francois, S. Wang, W. Bronzi, R. State, and T. Engel. Botcloud: Detecting botnets using mapreduce. In IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–6. IEEE, 2011.

[12] M. V. Gayoso, A´. F. Herna´ndez, and E. L. Herna´ndez. In Proceedings of the International Conference on Security and Management (SAM), pages 1–7. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2014.

[13] L. Zhuang, J. Dunagan, D. Simon, H. Wang, I. Osipkov, and D. Tygar. Characterizing botnets from email spam records. USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), Vol. 8: Pages 1–9, 2008.

[14] J. Chen, R. Fontugne, A. Kato, and K. Fukuda. Clustering spam campaigns with fuzzy hashing. In Proceedings of the Asian Internet Engineering Conference, ACM, 2014.

Volume: 6, Issue: 3,Special Issue: 3 ,Apr,2024 ,ISSN_NO: 2321-3337