IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EMAIL PHISHING DETECTION IN CYBERSECURITY

[1]Akshay Anand, [2]Prof Sowmyashree K.M
[1]Research Scholar, [2]Assistant Professor
[1]Department of MCA
[1]PES College of Engineering, Mandya, Karnataka, India

***Abstract*:** Generating a fraudulent communication that appears to come from a reputable source is known as phishing. The most widely used mode of communication is e-mail. The purpose is to steal personal information like credit card numbers and login credentials, or to infect the victim's machine with malicious software. Phishing is a common type of cyberattack that everyone should be aware of. Phishing starts with a phone call, e-mail, or other form of communication designed to entice a victim. The communication looks to come from a reliable source in this form of attack. If the attacker is deceiving the victim, it is usually advised to give crucial information to a fraudulent website. Malware can be downloaded on occasion downloaded to the machine that could be used as a target Attackers profit financially by collecting their victims' credit card information or other personal information. Phishing e-mails are occasionally sent to employees in order to get login information or other personal information in order to conduct a complex attack on a company. Phishing is a popular entry vector for cyber-attacks like Advanced Permanent Threats (APT) and ransomware. During a phishing assault, attackers can collect background information about the victim's personal and professional history, hobbies, and activities by using social engineering and other public information resources, such as social networks like LinkedIn, Facebook, and Twitter.

***Index Terms* — Phishing, Forensic analysis, NLP Algorithm, Random Forest, Cybersecurity.**

I. INTRODUCTION

Phishing is a type of social engineering that involves misleading users. Users are frequently enticed by communications ostensibly coming from reputable sources such as social media sites, shopping sites, banks, online payment processors, or IT administrators. Phishing is the deception of a trustworthy person in an electronic connection in order to get sensitive information such as usernames, passwords, and credit card numbers. It's usually done through email spoofing or instant messaging, and it often urges consumers to enter personal information on a fake website that looks and feels exactly like the real one.

A communication that seems to have been sent by a known contact or company is frequently sent to a victim. The assault is launched either through a malicious file attachment containing phishing malware or by links to malicious websites. In either case, the goal is to infect the user's device with malware or drive the victim to a dangerous website. to persuade individuals to reveal personal and financial information such as passwords, account IDs, or credit card numbers. Successful phishing messages, which are typically disguised as messages from a well-known firm, are difficult to differentiate from genuine ones: phishing emails might include corporate logos and other recognizable images, as well as data obtained from the misrepresented company. Malicious links in phishing communications are frequently crafted to look like they go to the faked company. Subdomains and misspellings, as well as other link manipulation strategies, are all typical tricks.

II. LITERATURE SURVEY

[1] The paper gives a brief overview of semantic analysis of text, the SEA Hound detection algorithm, and machine learning for blacklist generation, all of which are used to detect targeted phishing email attacks. The detection results are effective for e-mails that are entirely composed of text, and semantic information is a strong indicator of social engineering. [2] This paper explains how they specialize on email communication, which is the most common way for such assaults to be launched. The system is designed to detect phishing emails that do not contain any links and instead rely on the victim's curiosity to get them to respond with sensitive information. [3] This paper provides a brand Protects Anti-Phishing Email Analysis package, Spam Assassin, and Sophos Pure Message. These features help to reduce common subterfuges used in crafting phishing emails. [4] This survey gives a broader classification of defense mechanisms, and we provide a set of features used for phishing detection associated with these features ranked according to their ability to classify the phishing emails effectively.[5] Until recently, there hasn't been a singular answer to the problem of phishing. Browsers are made capable of setting up systems that identify and warn of potential phishing attempts in order to implement it. We looked at various technical phishing attempts that can originate from any type of communication medium, including email, instant messaging, websites, and social media. Suspicious texts, phishing-like URLs, phishing web sites, and

malicious PDF documents attached to those platforms were all taken into account. The methods of lexical analysis for suspicious texts and fraudulent URLs, as well as recent machine learning approaches and pattern recognition techniques for both fraudulent URLs and phishing web sites, were discussed.

III. PROPOSED SYSTEM

To detect phishing web sites, we offer a content-based approach. It's a project implementation in which our system crawls the source sites and retrieves all URLs and information. Users should be careful if they get an email with a link to a phishing attack. Then, using that URL as input, our system will crawl the link, retrieve all URLs, and compare them using a machine learning algorithm (NLP), attempting to determine if they are similar or not. The system will then determine who owns the URL's information. The information will be analyzed by the system, and the findings will be shown to the user.

- Improved accuracy and efficiency.
- The automated classifier process meets risk and compliance.
- Forensic investigation provides further information about the email vulnerability.
- The tool gives you more information and divides the components into actionable and non-actionable.

IV. SYSTEM ARCHITECTURE

The project system architecture is described in the diagram. It describes the system's flow. A phishing email is sent by the sender. The email will then move through an email server and arrive at the recipient's end. The letter will pass via a customized firewall before reaching the victim's inbox, where it will be analysed and scanned for any dangerous terms using Natural Language Processing. If phishing is found, the email is banned and categorized using the Random Forest algorithm and natural language processing (NLP).

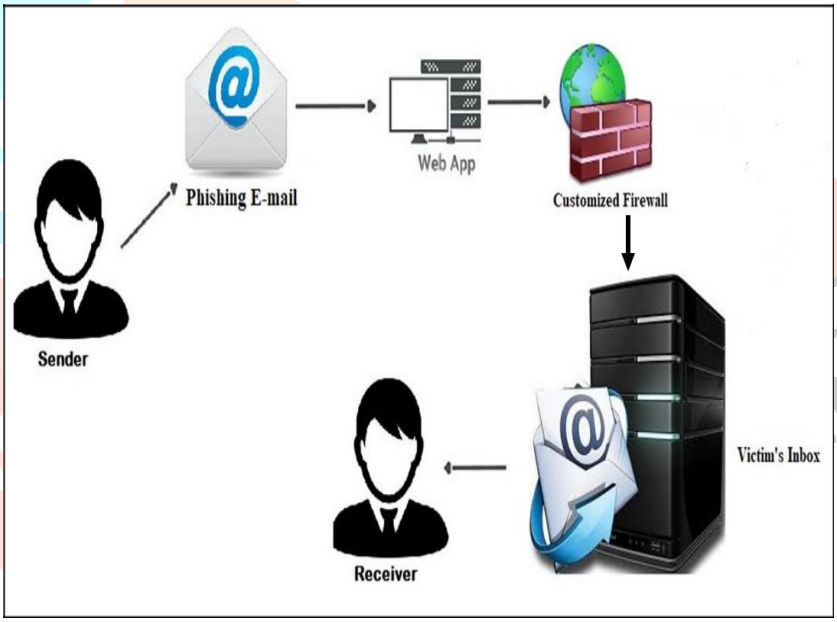

Fig 1: System Architecture of Proposed System

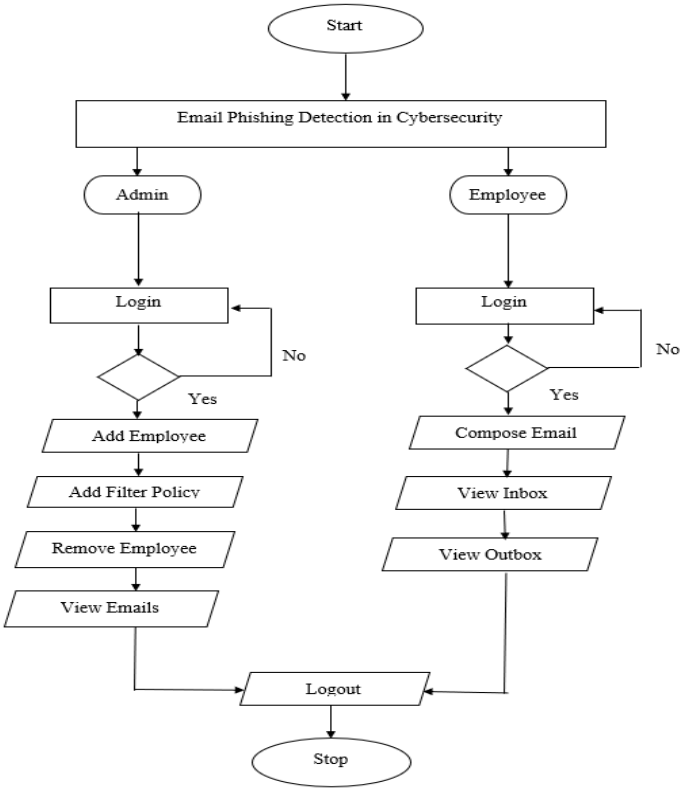

Fig 2: Workflow Diagram

V. IMPLEMENTATION

The project is implemented in modular approach. Each module is coded as per the requirements and tested and this process is iterated till the all the modules have been thoroughly implemented.

- Filter rule creation
- Application of NLP
- Email header analysis
- Auto suspension of malicious user
- Attachment detector and analysis
- Report generation for the threats in emails

Natural Language Processing:

Natural language processing methods enhance computers by emulating human language understanding. NLP analyses numerous components of human language, such as syntax, semantics, pragmatics, and morphology, to comprehend its structure and meaning. Automatic summarizing of the major points of a text or document is achieved using NLP algorithms. NLP algorithms are also used to organize the information, as well as in email routing and spam screening, and to classify text according to predetermined categories or classes.

NLP steps are as follows: -

a. Tokenization: In this phase sentences are divided into streams of individual tokens that are differentiated by spaces.

b. Stop Word Removal: Stop words are the words that occur frequently. These words must be eliminated because they influence the sentences that contain these words.

c. Case Folding: It is to reduce all letters to lower case.

d. Stemming: Stemming is the reduction of derived or inflected words to their stem, base, or root form, which is usually a written word form.

e. Keyword Extraction: It is a text analysis technique that extracts the most frequently used and important words and sentences from a document automatically.

Random Forest:

Random forests are widely used in data science contests and in real-world situations. They're usually accurate, don't require feature scaling or categorical feature encoding, and only require little parameter tweaking. They may also be simpler to learn than more complex models such as neural networks. Multiple random decision trees form a random forest. The trees have two different sorts of unpredictability. Each tree is first constructed using a random sampling of the original data. Second, a subset of features is chosen at random at each tree node to get the best split.

Steps: -

S1: Random sample selection from the dataset.

S2: After that, this algorithm will build a decision tree by defining the condition for each sample. The forecast result from each decision tree will then be obtained.

S3: Assign each data point to the closest cluster by calculating its distance from each centroid

S4: Calculate cluster centroids: The centroid of data points is computed by taking into account the 42 parameters.

S5: Assign each point to the cluster centroid nearest to it: Note that only the data point at the bottom is assigned to the cluster, despite the fact that it is closer to the cluster centroid. As a result, we place that data point in a cluster.
S6: Finally, choose the prediction result with the most votes as the final prediction result.

VI. EXPERIMENTAL RESULT

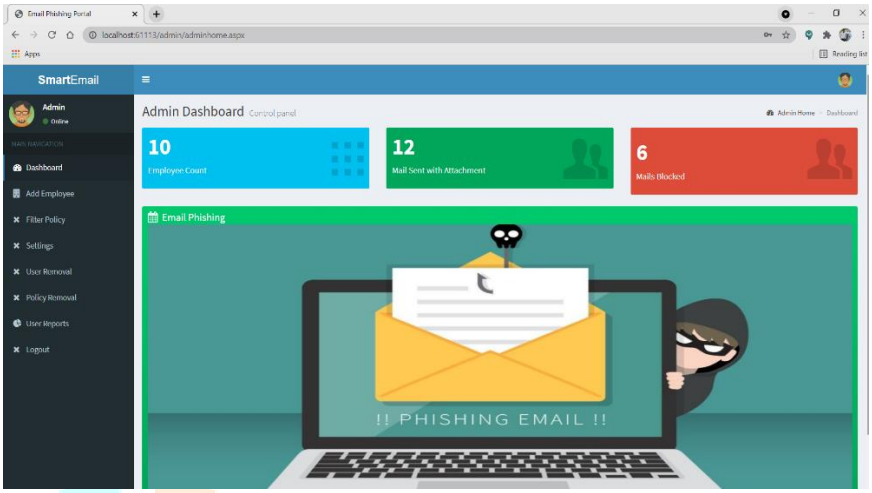

Fig 3: Admin Home Page

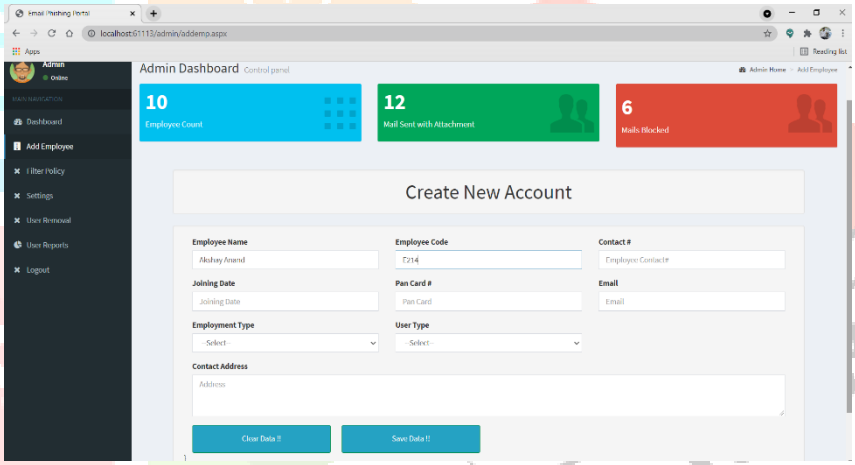

Fig 4: Adding New Employee

Here Admin going to add the Employee details to the System

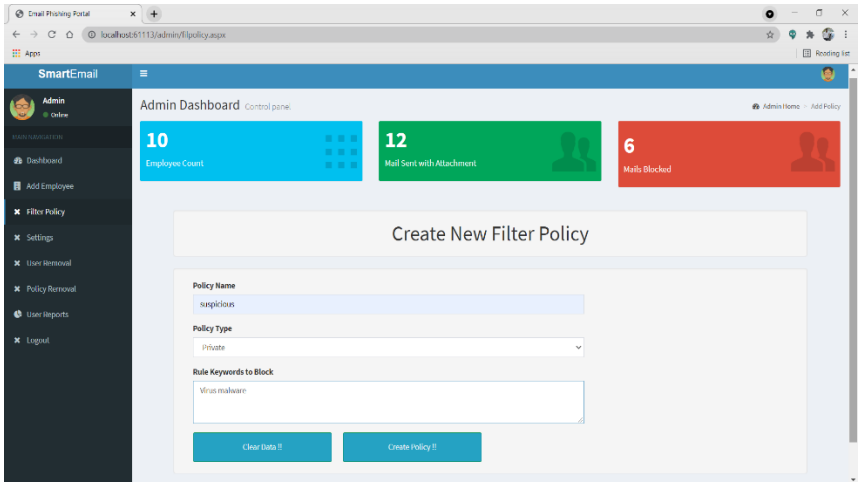

Fig 5: Creating New Filter Policy

Admin will create filter policy depending on company requirements to detect the phishing email or not

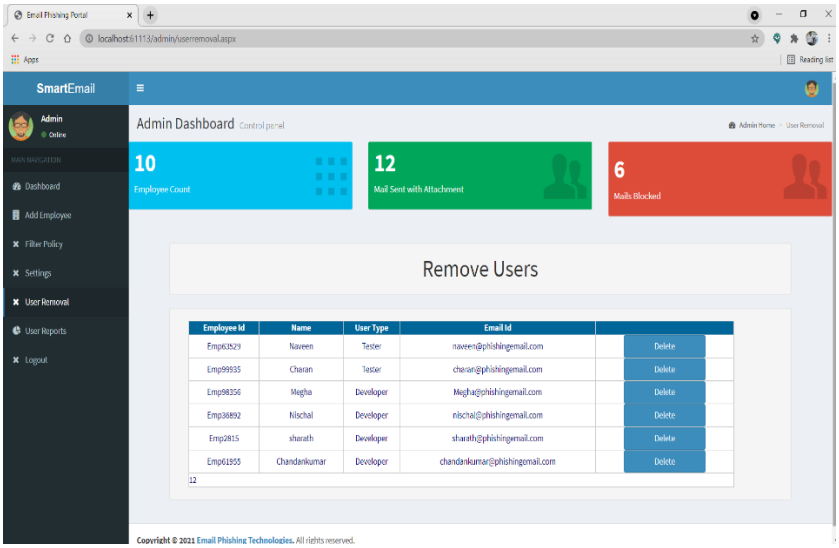

Fig 6: User Removal

The admin can remove users who are not working in the company

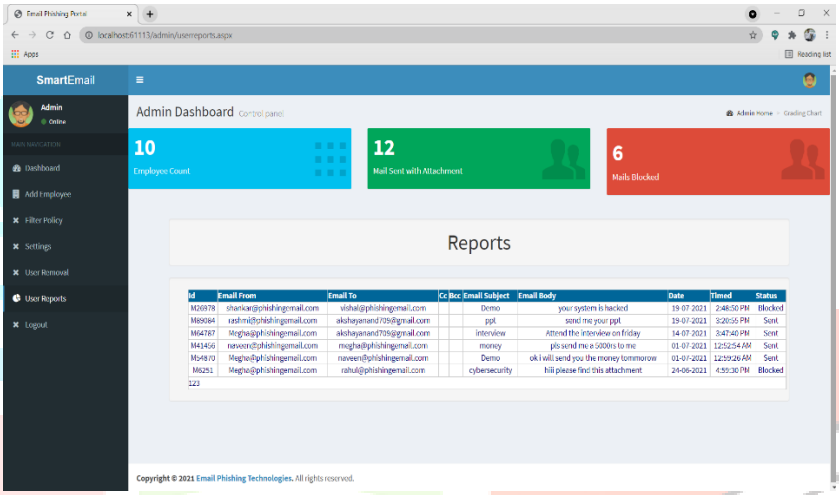

Fig 7: User Reports

Admin can view the reports of phishing mails blocked and non-phishing mails received

VI. CONCLUSION

In this project, phishing Using Natural Language Processing (NLP) approaches and Machine Learning algorithms such as Random Forest, this study detects phishing, which has become a serious network interconnection certainty difficulty generating financial losses of billions of dollars to enterprises. The metadata linked with emails and URLs implanted in phishing emails is analyzed, including sender and receiver mail ids, subject, body, and internet protocol address. The internet is used to control the physical components.

Random Forest, an anti-phishing algorithm based on the desired qualities, is then created. As a result of the Machine Learning Techniques performed in this work, phishing is a well-received kind of internet fraud that results in the revealing of financial and personal information. Some technical phishing attempts that come via communication media and suspicious texts are traversed in this study, as well as phishing like URLs and linked malicious PDF documents.

Several up-to-date and broad collective traits for detecting phishing attacks have been identified using Natural Language Processing. This method is based on textual analysis. As a result, it is effective at detecting phishing scams that use text messages.

VII. FUTURE ENHANCEMENT

In the future, the approach will be upgraded to tackle CSS (Cross Site Scripting) attacks from a greater distance. To identify a suspicious web page, more effective inferring criteria can be constructed, and tactics can be devised to determine if it is a phishing target. If Embedded Gadgets, such as voice messages, are present, the algorithms can be improved to detect phishing attacks.

REFERENCES

[1] Tianrui Peng, Ian G. Harris, Yuki Sawa" Detecting Phishing Attacks Using Natural Language Processing and Machine Learning, IEEE" 2018.

[2] Juan Chan, Chuanxiong Guo, —Online Detection and Prevention of Phishing Attacks,2006.

[3] Shivam Aggarwal, Vishal Kumar, S D Sudarsan, —Identification and Detection of Phishing Emails Using Natural Language Processing Techniques, IEEE Conference, 2016.

[4] Christopher N. Gutierrez, Taegyu Kim, Raffaele Della Corte, Jeffrey Avery, Dan Goldwasser, Marcello Cinque, Saurabh Bagchi, —Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks, 2016.

[5] Serafeltin Senturk, Ibrahim Sogukpinar" Email Phishing Detection and Prevention by Using Data Mining Techniques", 2017.

[6] Ebubekir Buber, Ozgur Koray Sahingoz, —Feature Selections for the Machine Learning based Detection of Phishing Websites, 2017.

[7] Naghmeh Moradpoor, Benjamin Clavie, Bill Buchanan, —Employing Machine Learning Techniques for Detection and Classification of Phishing Emails, 2017.

[8] Xi Zhang, Yu Zeng, Xiao-Bo Jin, Zhi-Wei Yan, Guan-Gang Gang —Boosting the Phishing Detection Performance by Semantic Analysis", 2017.

[9] Surbhi Gupta, Abhishek Singhal, Akanksha Kapoor, —A Literature Survey on Social Engineering Attacks: Phishing Attack, 2016

[10] Dhananjay Merat1, Anurag Patil, "A Machine Learning Approach for Phishing and Its Detection Techniques", 2020