Blockage of Phishing Using Machine Learning

P.Vignesh
Sri Sairam Institute of Technology

C.A.Sri Puvaanesh
Sri Sairam Institute of Technology

M. Bala Murugan
Sri Sairam Institute of Technology

S. Jacquilin Veda Jancy
Sri Sairam Institue of Technology .

***Abstract*— Trying to gather personal information through deceptive ways is becoming more common nowadays.In order to assist the user to be aware of the access to such websites, the implemented system notifies the user through emailand also pop-up, when trying to access a phishing site. This paper proposes an approach of phishing detection system to detect blacklisted URL also known as phishing websites, so that individual can be alerted while browsing or accessing a particular website. Therefore, it can be utilized for identification and authentication and become a legitimate tool to prevent an individual from getting tricked.**

Keywords: **LSTM, CNN, Machine Learning**

I. INTRODUCTION

Phishing can be defined as impersonating a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc. Phishing frauds might be the most widespread cybercrime used today. There are countless domains where phishing attack can occur like online payment sector, webmail, and financial institution, file hosting or cloud storage and many others. The webmail and online payment sector was embattled by phishing more than in any other industry sector. Phishing can be done through email phishing scams and spear phishing hence user should be aware of the consequences and should not give their 100 percent trust on common security application. MachineLearning is one of the efficient techniques to detect phishing as it removes drawback of existing approach.

The objectives which is the most vital thing in proposed project is to verity the validity of the website by capturing blacklisted URLs. To notify the user onblacklisted website through pop-up while they are trying to access and to notify the user on blacklisted website through email while they are trying to access. This proposed project will allow administrator to add blacklisted URL's in order to alert user during their inquiry.

The two scope of project, which is well known as user scope and system scope. User has some responsibility towards the system. The system includes a few standards and policies that requires to be obliged in order to comply the system. The user can be notified if blacklisted website is being accessed. The admin can capture the blacklisted URL's to alert user. The system involves features like capturing blacklisted website, viewing blacklisted website, displaying pop-up notification andalso displaying email notification.

II. LITERATURE SURVEY

The current situation that is majority of the population has been fooled into giving their personal details to hackers without noticing it. Many blacklisted website has been publish to appear as an original site in order to trap user by asking them to input their personal details. For example, password, bank account, email address and etc. Phishing activity in early 2016 was the highest ever recorded since it began monitoring in 2004.The total number of phishing attacks in 2016 was 1,220,523. Tis was a 65 percent increase over 2015. In the fourth quarter of 2004, there were 1,609 phishing attacks per month. In the fourth quarter of 2016, there was an average of 92,564 phishing attacks per month, an increase of 5,753% over twelve years. According to the Anti-Phishing Working Group (APWG), there are at least 47, 324 phishing attacks and a top-ten American bank estimates that at least US$300 is lost for every hour that a phishing site remains up. Machine learning is that the science of obtaining computers to act while not being expressly programmed. Machine Learning was implement to develop this proposed system. Machine learning techniques identifies phishing URLs typically assess a URL based on some feature or set of features extracted from it. .Thus, before coming to conclusion that this was the major problem, related products were examined and compared view their libation before progressing to the proposed project.

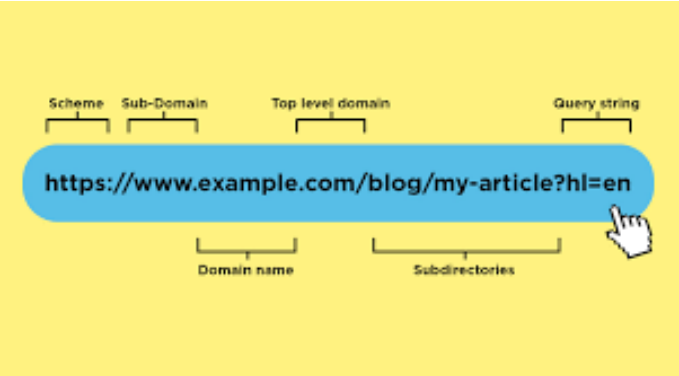
Phishtank was proposed to carry out the inspection once a link has been pasted on the section given. This allow user to keep on track of faked website. They can copy and paste the link in order to identify whether the site that they are going to access is safe or not safe. User can use the website search feature directly or they can use information from PhishTank through its API. A search engine displayed on PhishTank website is to be used as the first method. Using its API will be the second method. API service can be avail by software builder after registering themselves on PhishTank website. Both methods mentioned above do not cost a single penny. The purpose of API's usage is for user who has basis information on software development. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website.

PhishZoo was proposed to evaluate a new method for web phishing detection based on profiles of complex sites' appearance and content. PhishZoo makes profiles of sites comprising of the website contents and images displayed. These profiles are kept in a local folder and are either synchronized against the newly loaded

sites at the time of loading or against risky sites for instance, links in email offline. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website.

GoldPhish was proposed to perceive and report phishing sites. This was done by using optical character recognition (OCR) to recite the text from an image of the page precisely from the company logo, grasping the top hierarchical areas from a search engine, and comparing them with the current web site. The forte of the tool lies in the user's capability to recognize famous company logos. A phishing site cannot change a familiar company logo without the phishing target perceiving. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website.

III. PROPOSED SOLUTION



Detecting phishing URLs using machine learning involves training a model to distinguish between legitimate and malicious URLs based on various features. Here's a proposed solution:

1. Data Collection:
- Gather a diverse dataset of both legitimate and phishing URLs.
- Include features such as URL length, domain age, presence of HTTPS, use of subdomains, etc.
- Ensure a balanced dataset to avoid bias.

2. Data Preprocessing:
- Clean and preprocess the data, handling missing values and normalizing numerical features.
- Extract relevant information from URLs, such as domain and path.

3. Feature Extraction:
- Extract features that are indicative of phishing, e.g., the presence of suspicious keywords, IP address in URL, or the use of redirects.

4. Model Selection:
- Choose a suitable machine learning model. Decision trees, random forests, and gradient boosting classifiers often work well for this task.
- Consider ensemble methods for improved accuracy.

5. Training the Model:
- Split the dataset into training and testing sets.
- Train the model using the training set and validate its performance on the testing set.
- Fine-tune hyperparameters for optimal performance.

6. Evaluation:
- Evaluate the model using metrics such as accuracy, precision, recall, and F1 score.
- Use techniques like cross-validation to ensure robustness.

7. Integration:
- Implement the trained model into the phishing detection system.
- Regularly update the model with new data to improve its effectiveness.

8. Real-time Monitoring:
- Implement real-time monitoring to continuously assess the performance of the model.
- Integrate feedback mechanisms to adapt to emerging phishing techniques.

9. User Education:
- Educate users about recognizing phishing attempts and encourage safe browsing habits.

10. Collaboration:
- Collaborate with cybersecurity communities to stay informed about the latest phishing threats.
- Share insights and data for collective improvement.

11. Adaptive Learning:
- Implement techniques for adaptive learning to keep the model updated and effective against evolving phishing strategies.

12. Post-Deployment Analysis:
- Monitor the model's performance post-deployment and address any issues promptly.
- Collect feedback from users and security analysts for continuous improvement.

METHODOLOGY:

The proposed solution of phishing has four stages :

1. Input URL
2. Data Pre-Processing
3. Detection using DL models[CNN,LSTM,CNN-LSTM]
4. URL Classification – Phishing URL

   Legitimate URL

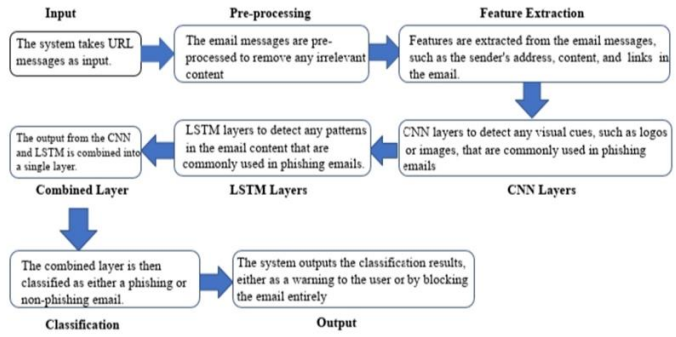
CNN and LSTM are two types of neural networks that can be used in phishing prevention.

Convolutional Neural Networks (CNN) are commonly used in image and speech recognition tasks. However, they can also be used to analyze the textual content of phishing emails or websites. CNNs can extract relevant features from the text of phishing messages, such as URLs and email addresses, to detect and classify them as phishing attempts. By training the CNN on a large dataset of known phishing attempts, it can learn to recognize and detect new phishing messages with a high degree of accuracy.

Long Short-Term Memory Networks (LSTM) are a type of recurrent neural network that can be used to analyze sequences of data, such as the textual content of emails or websites. LSTMs can learn to identify patterns in sequences of data and use them to predict whether a new sequence of data is a phishing attempt. By training the LSTM on a large dataset of known phishing attempts, it can learn to recognize and detect new phishing messages with a high degree of accuracy.

By combining CNN and LSTM, researchers have developed more advanced machine learning models for phishing prevention. These models can analyze both the textual and visual content of phishing messages to identify patterns and detect new phishing attempts. With the increasing sophistication of phishing attacks, these advanced machine learning models are becoming more important for protecting users from phishing attemp

Block Diagram:



IV. CONCLUSION

In conclusion, this system is designed for resources are used as intended, prevents from valuable information from leaks out, produce better control mechanism and alerts the user to keep their private information safe. Like any other programs, there are improvements which could be made into this system. Based on the capabilities which the current system processes, text message integration would a great recommendation that could be made to improve the program in the future. The future version of the application could also implement an option to directly notify the blacklisted website with a text message. The program could be made to access the list as an attachment. This text message integration function would further the usability of the application.

V. REFERENCES

[1] Matthew Dunlop, Stephen Groat, David Shelly (2010) " GoldPhish: Using Images for Content-Based Phishing Analysis"

[2] Rishikesh Mahajan (2018) "Phishing Website Detection using Machine Learning Algorithms"

[3] Purvi Pujara, M. B.Chaudhari (2018) "Phishing Website Detection using Machine Learning : A Review"

[4] David G. Dobolyi, Ahmed Abbasi (2016) "PhishMonger: A Free and Open Source Public Archive of Real-World Phishing Websites"

[5] Satish.S, Suresh Babu.K (2013) "Phishing Websites Detection Based On Web Source Code And Url In The Webpage"

[6] Purvi Pujara, M. B.Chaudhari (2018) "Phishing Website Detection using Machine Learning : A Review"

[7] Satish.S, Suresh Babu.K (2013) "Phishing Websites Detection Based On Web Source Code And Url In The Webpag

[8] Tenzin Dakpa, Peter Augustine (2017) "Study of Phishing Attacks and Preventions"

[9] Ping Yi (2018) “Web Phishing Detection Using a Deep Learning Framework”

[10] Jalil Nourmohammadi Khiarak (2017) “What is Machine Learning”

[11] Sadia Afroz, Rachel Greenstadt (2018) “PhishZoo: An Automated Web Phishing Detection Approach Based on Profiling and Fuzzy Matching”

[12] Arun Kulkarni, Leonard L. Brown (2019) “Phishing Websites Detection using Machine Learning”

[13] Rohan Saraf , Mayur Khatri , Mona Mulchandani (2014) “Phish Tank-A Phishing Detection Tool”

[14] Sadia Afroz, Rachel Greenstadt (2017) “PhishZoo: Detecting Phishing Websites By Looking at Them”

[15] Matthew Dunlop, Stephen Groat, David Shelly (2010) " GoldPhish: Using Images for Content-Based Phishing Analysis”