International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 588-593 www.ijaem.net ISSN: 2395-5252

Detection of Phishing Website

1st Sai Charan Soma, 2nd Vavilapalli Sindhu, 3rd T. Leela Krishna, 4th B.V. Hiranmayee Satya, 5th G. Taraka Rama Rao

Dept of Computer Science and Engineering GITAM (Deemed to be University)Visakhapatnam, India
Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India
Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India
Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India
Assistant Professor, Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India

Date of Submission: 05-04-2023 Date of Acceptance: 15-04-2023

ABSTRACT—Phishing is a sort of online fraud in which attackers pretend to be a reputable website or institution in order to obtain sensitive information from unwitting victims. The goal of phishing website URLs is to steal personal information such as user names, passwords, and online banking activities. Phishers employ webpages that are aesthetically and semantically identical to legitimate websites. As technology advances, phishing attempts get more sophisticated, which must be avoided by employing anti-phishing systems to identify phishing. To combat this problem, researchers have developed various techniques for detecting phishing websites. This project aims to create a phishing website detection system that utilizes machine learning algorithms to classify websites as legitimate or phishing. The system will be trained on a dataset of known phishing and legitimate websites and will use features such as website structure, URL patterns, and content to make predictions. The system will then be evaluated using standard metrics such as accuracy, precision, and recall. Ultimately, the purpose of this project is to develop a tool that may assist internet users in avoiding phishing fraud

I. INTRODUCTION

Phishing attacks are among online systems and networks' most significant security threats. These assaults often include bogus websites that seem like real websites to fool users into submitting personal information such as usernames, passwords, and credit card information. Phishing attacks can lead to significant financial losses, reputational damage, and legal liabilities for individuals and organizations.

To address this threat, developing a phishing website detection system has become a critical need. A phishing website detection system is a software solution to identify and block fraudulent websites, protecting users from phishing attacks. The technology analyses website elements and classifies them as authentic or phishing using machine learning techniques.

The need for a phishing website detection system has become more significant with the increasing sophistication of phishing attacks. In the past, phishing websites were relatively easy to detect, as they were often poorly designed and had obvious spelling and grammatical errors. However, modern phishing websites are much more sophisticated, often using advanced techniques such as URL obfuscation and social engineering to trick users into providing sensitive information.

We propose developing a phishing website detection system that can accurately classify websites and provide reliable protection against phishing attacks to address this challenge. The system will use machine learning algorithms to analyze website features such as content, URL structure, and HTML tags to identify and block fraudulent websites. The system will also use a continuously updated database of known phishing websites to improve its accuracy and effectiveness.

The development of a phishing website detection system has several benefits. First and foremost, it can significantly improve the security of online systems and networks by preventing users from accessing fraudulent websites and providing sensitive information. This can help to avoid financial losses and reputational damage caused by phishing attacks. Developing a phishing website detection system can also enhance user confidence in online systems and networks, leading to increased adoption and usage.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 588-593 www.ijaem.net ISSN: 2395-5252

PHISHING

Phishing is an online scam in which users are duped into disclosing sensitive data such as passwords, usernames, credit card numbers, and other personal information via bogus emails, websites, and other means of electronic contact.

Phishing scams are typically designed to look like legitimate requests for information from a trusted source, such as a bank or an online service provider. The scammer's goal is to trick the victim into clicking on a link or providing personal information that can be used for fraudulent purposes, such as identity theft or financial fraud.

Phishing schemes may take numerous forms, including email, instant messages, social media posts, and fake websites. In some cases, phishing emails or messages may contain malware or other malicious software that can infect the victim's computer or mobile device. To avoid phishing scams, be skeptical of unsolicited emails or texts, and never click on links or disclose personal information unless you are confident of the source.

Phishing can have several adverse effects on both individuals and organizations, including:

Financial loss: Financial information, such as credit card numbers and bank account information, is frequently stolen through phishing schemes. Victims may incur financial losses or even identity theft if this information gets into the wrong hands.

Damage to reputation: Whenever a person or business falls victim to a phishing scam, it can harm its reputation and lose consumer or partner confidence.

Loss of data: Besides financial information, phishing scams may also be used to steal sensitive data such as login credentials or intellectual property. This can lead to a loss of competitive advantage and may even result in legal consequences.

Disruption of services: Phishing attacks can also disrupt services or operations, mainly if malware or other malicious software is introduced into an organization's network.

Psychological impact: Victims of phishing scams may also suffer from psychological discomforts, such as worry and stress, which can hurt their general well-being and productivity.

Overall, phishing can significantly impact individuals and organizations, and protecting against these attacks is essential.

II. RELATED WORKS

There have been several related works in the area of phishing website detection. Some notable examples include:

[1]"A Hybrid Machine Learning-Based Phishing Detection System" by Khan et al. (2019): This paper proposes a hybrid machine learning-based phishing detection system that combines several machine learning algorithms to improve the accuracy of phishing detection.

[2]"A Deep Learning Approach for Phishing Websites Detection" by Murgante et al. (2020): This paper presents a deep learning approach for phishing website detection that uses convolutional neural networks (CNNs) to analyze website content and detect phishing attacks.

[3]"A Novel Approach to Detect Phishing Websites using Machine Learning Techniques" by Patel et al. (2020): This paper proposes a novel approach to detect phishing websites using machine learning techniques such as K-means clustering and random forest.

[4] "An Intelligent Anti-Phishing System based on Machine Learning" by Li et al. (2019): This paper proposes an intelligent anti-phishing system based on machine learning that uses decision tree algorithms to detect phishing websites. [5]"Phishing Detection using Machine Learning: A Review" by Mehta et al. (2019): This paper provides a comprehensive review of phishing detection using machine learning and highlights the strengths and weaknesses of different approaches.

Overall, these related works demonstrate the potential of machine learning and other relevant techniques to improve the accuracy and effectiveness of phishing website detection, and they provide valuable insights into the state-of-the-art in this area.

III. PROPOSED SYSTEM METHODOLOGY

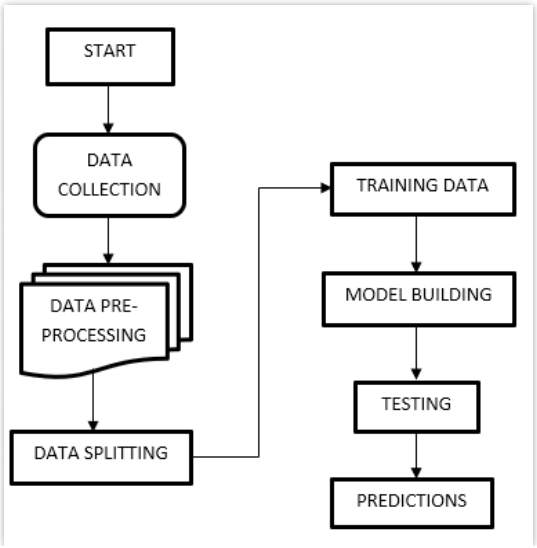

Fig. 1: Workflow

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 588-593 www.ijaem.net ISSN: 2395-5252

The above workflow are the steps followed for creating the model. It consists of three main parts. Firstly, we collect the data and pre-process the data, after pre- processing we split the data into training and testing data. Now the training data is used for model building and hence the model is built. This is the second and most important part of the workflow. Lastly to evaluate the model the testing dataset is used and hence required predictions are made.

To elaborate each part of the workflow is explained below:

Data collection: The primary step is to collect a dataset of legitimate and phishing websites. This dataset is used to train and test the machine-learning algorithms that will be used for detection.

Data pre-processing: The extracted features are pre-processed to prepare them for machine learning algorithms. This can involve scaling, normalization, and other techniques to reduce noise and improve accuracy.

Data Splitting: According to the features present in the dataset we opt to split our data into 70% to train the model and remaining 30% for testing the model.

Training Data: The pre-processed dataset trains the machine learning algorithms. Different types of algorithms can be used, including decision trees, random forests, etc.

Model Building:The trained model uses a test dataset to measure its accuracy and precision. The model can be further refined and optimized based on these evaluation metrics.

Testing:Once the model is trained and tested, it can be deployed in a production environment for real-time phishing detection.

Predictions: The models have given appropriate predictions and the predictions were made on testing data which has shown high accuracy matching the state of art accuracy values.

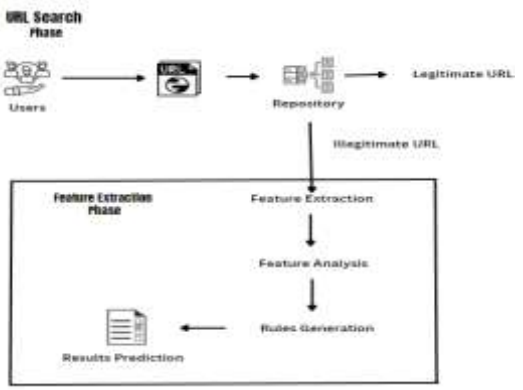

Fig.2: Architecture

From the fig.2 we can demonstrate how our model actually works in real time applications. Initially the users provide website URL to check whether there is a possibility of phishing or not. The URL is provided as an input then the model goes through the repository, if the provided URL is already classified as legitimate in the repository the model simply classifies the URL as legitimate and provide user that the URL is safe. In contrary if the URL is not available the feature extraction phase begins. In this phase the features of the URL are extracted, those features include domain length, URL length, URL depth, and domain age including with its expiry. These features are now analyzed in feature analysis phase. In Rules generation phase the features which are analyzed are compared and if the URL is matching with existing legitimate features the features are given to model for training and if the URL is illegitimate those features are also fed to the model and provides user that the URL is Illegitimate. As we are using machine-learning algorithms, in both the cases the features are given to the model so that the model learns from the user inputs as well and provide more accurate results on further basis.

IV. IMPLEMENTATION

Data gathering, data preprocessing, feature engineering, model training, and model deployment are some of the phases in the implementation of the project to detection of phishing websites. The initial stage is to gather data from diverse sources, including governmental organizations, academic journals, etc. Data on phishing websites, cyber-attacks like phishing, blue jacking and other existed data should all be included. Every project involving data analysis or machine learning, such as one involving the phishing attack, must first undergo data preparation. It entails modifying, cleaning, and preparing the raw data so that it is appropriate for modelling and analysis. Data cleaning include deleting or adding missing numbers, fixing mistakes, and dealing with outliers in the data. Use of an effective algorithm to detection of phishing website andprovide recommendations whether the website is legitimate or illegitimate.

A. Data set interpretation

The data set named" Detection of Phishing Website" is sourced from Kaggle, a popular platformfor data science enthusiasts. This data set contains informationon many websites URL's including some of their features which helps to train the model. The data set includes the followingcolumns: Domain, URL Length, URL depth, Domain Age, Domain End, Web Traffic, etc.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 588-593 www.ijaem.net ISSN: 2395-5252

B. Data set processing:

Processing the data set is an essential step in getting the information ready for analysis and machine learning. It requires a number of procedures, including data cleansing, data transformation, and feature engineering. These are some crucial factors to think about while processing data sets: Cleaning the data include eliminating duplicates, fixing data mistakes, dealing with outliers, and deleting or imputed missing numbers.

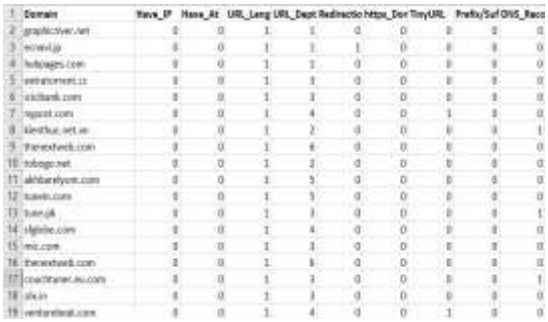

Fig.3: Pre-processed Data

C. Model Building:

Building a machine learning model that can make predictions or categorize data entails training a model using the prepared dataset. While training a model, keep the following points in mind: Selecting the optimal machine learning algorithm that is suitable for the dataset's nature and the job at hand is the first step. For diverse sorts of issues, including regression, classification, clustering, and association rule learning, there are several techniques available. Before training the model, hyperparameters must be specified for the majority of machine learning algorithms. To maximize the model's performance, certain hyperparameters must be tuned.

V.EVALUATION

Home Page: Here users view the home page of the Detection of Phishing web application.

Registration Page: Here users can register on the website by providing their details.

Login Page: Here users can log in to the website by providing their login credentials. If the user is new to the website, they must follow the above step.

Model: Here we can train our data using different algorithm.

Here the user can select particular algorithm to find the accuracy.

Fig.4: Model Selection

We have considered a pre-processed data as a testing data to train and test machine learning models to get the prediction values accurately.

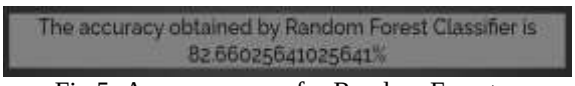

Fig.5: Accuracy score for Random Forest

By considering the processed data and Random Forest Classifier we have obtained the accuracy of 82.66025641025641% which is 83% approximately.

Fig.6: Accuracy score for AdaBoost

By considering the processed data and AdaBoost Classifier we have obtained the accuracy of 78.5% which is 79% approximately.

Fig.7: Accuracy score for XGBoost

By considering the processed data and XGBoost Classifier we have obtained the accuracy of 82.40384615384615% which is 82% approximately.

Prediction: This page shows the detection result that whether the website is a phishing website or legitimate.

Logout: Here the users can logout from the web application.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 588-593 www.ijaem.net ISSN: 2395-5252

S. No	INPUT	ACTUAL OUTPUT	EXPECTED OUTPUT	REMARKS
1.	https://www.gitam.edu/	Website is Legitimate	Website is Legitimate	Pass
2.	https://www.amazon.in/	Website is Legitimate	Website is Legitimate	Pass
3.	https://melissa.com	Website is illegitimate	Website is illegitimate	Pass
4.	http://east.dpsbangalore.edu.in/nice-admissions/	Website is illegitimate	Website is illegitimate	Pass
5.	https://uidai.gov.in/	Website is Legitimate	Website is Legitimate	Pass
6.	https://www.kvms.org.in/	Website is illegitimate	Website is illegitimate	Pass
7.	https://www.hotstar.com/in	Website is Legitimate	Website is Legitimate	Pass

Fig. 8: Testcases Table

We have tested the model using some testcases as mentioned in fig.5. So, in all the considered testcases our model has given very accurate results. In all the testcases which are mentioned, some of them may not be recognized by the users, since they were familiar with common websites that we use regularly. In this type of cases our model is very useful and helps users to detect the URL is legitimate or not.

VI. RESULTS

Fig.9. Predicts website is legitimate

Fig.10. Predicts website is illegitimate

By comparing the features and attributes in the processed data, the model can identify whether the provided URL was legitimate or illegitimate as each URL contains similar features in the dataset.

Algorithms Used	Accuracy(approx.)
Random Forest Classifier	83%
Ada Booster	79%
XG Booster	82%

Table: Accuracy score obtained w.r.t algorithms used

So, as the main purpose of the detection of phishing websites project is to find whether the URL of a particular website has a chance for phishing attack or not. With the help of processed data set we can train the model along with machine learning algorithms to predict the accurate results which can satisfy our requirement.

VII. CONCLUSION & FUTURE SCOPE

In conclusion, developing a phishing website detection system is crucial to ensuring the security of online systems and networks. Phishing attacks are a significant threat that can lead to financial losses, reputational damage, and legal liabilities for individuals and organizations. Therefore, a reliable and accurate phishing website detection system is needed to identify and block fraudulent websites.

The proposed approach for developing a phishing website detection system involves using machine learning algorithms to analyze website features such as content, URL structure, and HTML tags. This approach can accurately classify websites and provide reliable protection against phishing attacks. Additionally, the system will use a continuously updated database of known phishing websites to improve its accuracy and effectiveness.

The literature survey identified various approaches and techniques used in phishing website detection, including machine learning techniques, anomaly-based detection, feature selection, and neural network-based methods. These works provide valuable insights into developing an effective phishing website detection system.

The future scope of the detection of a phishing website project is vast, and there are several areas where it can be further developed and improved. Some potential areas of future research and development include:

Real-time monitoring: The system can be further developed to monitor websites in real-time and detect phishing attacks as soon as they occur. This would enable a more proactive response to phishing attacks and reduce the risk of successful attacks.

1.Continuous learning and adaptation: The system can be designed to continuously learn and adapt to new types of phishing attacks and update its detection algorithms accordingly. This would ensure that the system remains practical and up-to-date against the latest threats.
2.Mobile device support: The system can be optimized to detect phishing attacks on mobile devices, where users are increasingly accessing online services and conducting financial transactions.
In summary, the detection of phishing website projects has significant future scope for further research and development, and there are many areas where it can be improved to provide more robust and reliable protection against phishing attacks.

REFERENCES

[1]. Machado, L. G., dos Santos, J. A., & de Castro, A. L. (2017). A review of techniques for website classification. Journal of Network and Computer Applications, 94, 84-97.
[2]. Al-Rubaie, A. (2017). Phishing detection techniques: A systematic review. Journal of Network and Computer Applications, 88, 23-40.
[3]. Dalvi, N., Dhar, V., & Suciu, D. (2004). Mining the web for spam: A comparison of algorithms. Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, 535-540.
[4]. Mestre, J., & Krishnan, H. (2008). Phishing detection: A machine learning approach. Proceedings of the 2008 ACM symposium on Applied computing, 1519-1523.
[5]. Kumar, A., & Singh, M. (2010). A review on phishing detection techniques and challenges. International Journal of Computer Science and Information Security, 8(5), 128-136.
[6]. Pal, S., & Sahoo, S. (2017). A review on phishing website detection and prevention. International Journal of Innovative Research in Science, Engineering and Technology, 6(11), 6201-6210.
[7]. Fette, I., & Sompel, H. V. (2007). Phishing detection using open-source machine learning. Proceedings of the 2007 ACM workshop on Recurring malcode, 37-45.
[8]. J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.
[9]. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
[10]. T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.
[11]. M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
[12]. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.