ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 18, Issuse.1 March 2024

DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

Devarapalli srujana[1],Nagalla VenkataLakshmi[2],Vinay Podicheti[3], Sashirekha Ravula[4]

[1,2,3]B.Tech Student, Department of CSE (Data Science), Malla Reddy College of Engineering and Technology, Hyderabad, India.

[4]Assistant Professor, Department of CSE (Data Science), Malla Reddy College of Engineering and Technology, Hyderabad, India.

ABSTRACT:

The rapid growth of internet usage has led to an alarming increase in phishing attacks, which aim to deceive users into divulging sensitive information. While by clicking the fake URL's pop ups, we get hacked by the attackers and snippers. Malicious website is common and serious threat to cybersecurity. Machine learning techniques have emerged as a promising solution to this problem.In any phishing attack, the user can trapped into clicking some link to phishing website where user can reveal their sensitive information like username, password, mobile number, email address, etc.These features include textual content analysis, URL characteristics, SSL certificate properties, and behavioral patterns of users interacting with websites. The collected data is preprocessed and used to train a machine learning model, such as Random Forest, Gradient Boosting, or Deep Neural Networks, depending on the dataset's size and complexity.By continuously updating the model with new data, our approach ensures adaptability to emerging phishing threats.Additionally,the use of machine learning significantly reduces false positives and provides a more robust defense against phishing attacks.

Keywords-Phishing attacks,Feature extraction,Machinelearning,Accuracy,Real-time dataset.

I.INTRODUCTION:

This introduction aims to explore the application of machine learning algorithms in the detection of phishing websites.We will delve into the key challenges posed by phishing attacks, the role of machine learning in addressing these challenges, and the benefits it brings to the realm of cybersecurity.Phishing website detection is a persistent security threat, and this research compares classic supervised machine learning algorithms on all publicly available phishing datasets to distinguish the best performing algorithm, with ensembles and neural networks outperforming other classical algorithms.Hackers install malicious software on computers to steal

credentials, often using systems to intercept username and passwords of consumers' online accounts. Phishers use multiple methods, including email, Uniform Resource Locators (URL), instant messages, forum postings, telephone calls, and text messages to steal user information.The primary objective of phishing is to gain certain personal information for financial gain or use of identity theft. Phishing attacks are causing severe economic damage around the world. Moreover, Most phishing attacks target financial/payment institutions and webmail, according to the Anti-Phishing Working Group (APWG) latest Phishing pattern studies.In a recent period, attackers and hackers have an advantage using the Internet as their channel to attack the client or users device to gain information using fake URLs. So the users need a method to identify malicious URLs and help them from being victims of scam. So, in this project, we developed the method to identify the malicious and fake URLs with the help of Machine Learning. With Machine Learning algorithms it is possible to teach the machines, to identify the malicious URLs automatically. The project's goal is to detect phishing URLs and narrow down the best machine learning algorithm by evaluating each algorithm's accuracy rate, false positive rate, and false negative rate.Links to malware-infected websites may be included in phishing emails. Phishing is a type of social engineering method that takes advantage of flaws in current web security to deceive consumers. Legislation, user training, public awareness, and technical security measures are all being used to combat the rising number of reported phishing instances. In an age where phishing attacks continue to evolve in sophistication and scale, harnessing the capabilities of machine learning to proactively identify and thwart these threats is not only timely but also essential for safe guarding the digital ecosystem. This exploration into the realm of phishing detection using machine learning will provide valuable insights into how modern technology can be harnessed to protect individuals and organizations from falling victim to these deceptive online schemes. Uniform Resource Locators (URLs), sometimes known as "Weblinks," are the primary means by which users locate resources on the Internet.Our goal is to detect malicious Web sites are safe or malicious using the URLs. .Our experimental results demonstrate the effectiveness of the proposed approach in accurately identifying phishing websites. The machine learning model's ability to generalize and adapt to new phishing

techniques allows it to outperform traditional rule-based systems. Moreover, we employ various evaluation metrics, such as precision, recall, and F1-score, to assess the model's performance comprehensively.By continuously updating the model with new data, our approach ensures adaptability to emerging phishing threats. Additionally, the use of machine learning significantly reduces false positives and provides a more robust defense against phishing attacks. Online banking, credit card payments and debit card payments are also become quite popular since past few years. In this paper, we leverage a diverse set of features extracted from website content, structure, and network behavior.

II.LITERATURE REVIEW

In this literature review, we'll explore some key studies and developments in this field, highlighting the various techniques, datasets, and algorithms used for effective phishing detection.

According to H. Huang et al., (2009) proposed the frameworks that distinguish the phishing utilizing page section similitude that breaks down universal resource locator tokens to create forecast preciseness phishing pages normally keep its CSS vogue like their objective pages. S. Marchal et al., (2017) proposed this technique to differentiate Phishing website depends on the examination of authentic site server log knowledge. An application Off-theHook application or identification of phishing website. Free, displays a couple of outstanding properties together with high preciseness, whole autonomy, and nice language-freedom, speed of selection, flexibility to dynamic phish and flexibility to advancement in phishing ways. Fadi Thabtah et al. experimentally compared large numbers of ML techniques on real phishing datasets and with respect to different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more appropriate as anti-phishing solutions. Muhemmet Baykara et al. proposed an application which is known as Anti Phishing Simulator, it gives information about the detection problem of phishing and how to detect phishing emails. Spam emails are added to the database by Bayesian algorithm. Phishing attackers use JavaScript to place a legitimate URL of the URL onto the browsers address bar. The recommended approach in the study is to

use the text of the e-mail as a keyword only to perform complex word processing.

III.METHADOLOGY

The methodology for detecting phishing websites using machine learning algorithms typically involves several key steps. Below is an outline of a common approach to building a phishing detection system:

1. Data Collection:

- Gather a labeled dataset that includes examples of both phishing and legitimate websites. You can obtain these labels from reputable sources, such as PhishTank or by using web crawling techniques to identify and label phishing websites.

2. Data Preprocessing:

- Prepare the dataset by cleaning and preprocessing the data. This may involve removing duplicates, handling missing values, and converting data into a suitable format for machine learning algorithms. Additionally, you may need to extract relevant features from the URLs or web page content.

3. Feature Extraction:

- Extract meaningful features from the data that can be used as input for the machine learning model. Features can include URL-based features (e.g., domain length, presence of hyphens), content-based features (e.g., HTML tags, text analysis), and behavioral features (e.g., user interactions).

4.Dataset Splitting:

- Split the dataset into training, validation, and testing sets. The training set is used to train the machine learning model, the validation set helps in tuning hyperparameters and preventing overfitting, and the testing set is used to evaluate the model's performance.

5. Model Selection:

- Choose an appropriate machine learning algorithm or ensemble of algorithms for phishing detection. Common choices include decision trees, random forests, support vector machines (SVM), logistic regression, neural networks, and gradient boosting.

6. Model Training:

- Train the selected machine learning model(s) on the training dataset using the extracted features. Optimize hyperparameters to achieve the best performance. You may also consider techniques like cross-validation to assess model robustness.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 18, Issuse.1 March 2024

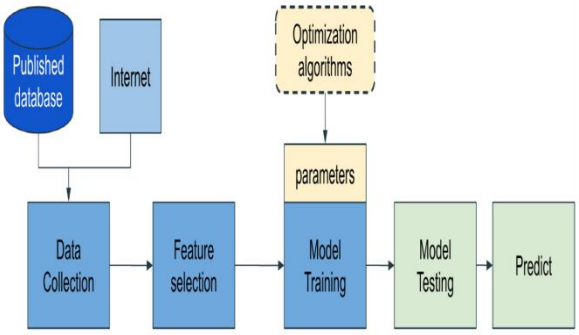
7. Model Evaluation:

- Evaluate the trained model(s) on the validation set and fine-tune the model if necessary. Common evaluation metrics for phishing detection include accuracy, precision, recall, F1-score, and ROC-AUC.

8. Testing and Deployment:

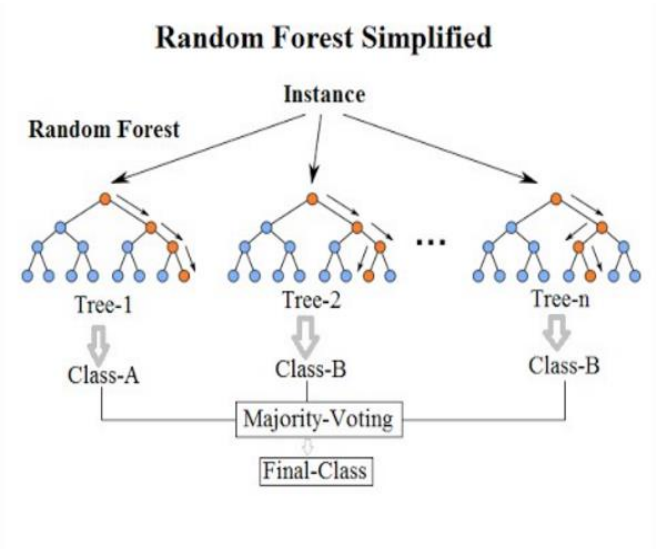
- Once you are satisfied with the model's performance on the validation set, assess its performance on the testing dataset, which simulates real-world conditions. If the results meet your criteria, proceed to deploy the model in a production environment.

It's important to note that the effectiveness of a phishing detection system depends not only on the machine learning algorithm but also on the quality and representativeness of the dataset and the continuous monitoring and maintenance of the system in response to evolving phishing threats.



IV.ALGORITHM

In our project, we have used machine learning algorithm called RFA(RandomForest Algorithm). This is a supervised learning algorithm which is used for both Classification and Regression techniques used in ML. The concept of the randomforest algorithm is ensemble learning which means combining the multiple classifiers to solve the complex problem and improve the performance. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees[citation needed]. However, data characteristics can affect their performance.



V.IMPLEMENTATION

To detect URLs within a given text using Python and regular expressions, we can follow a straightforward process. First,

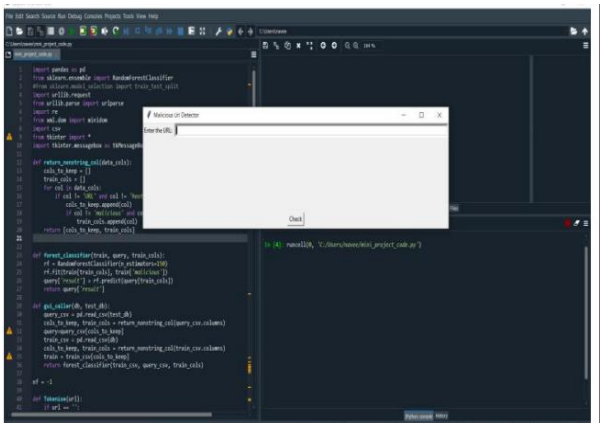
INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 18, Issuse.1 March 2024

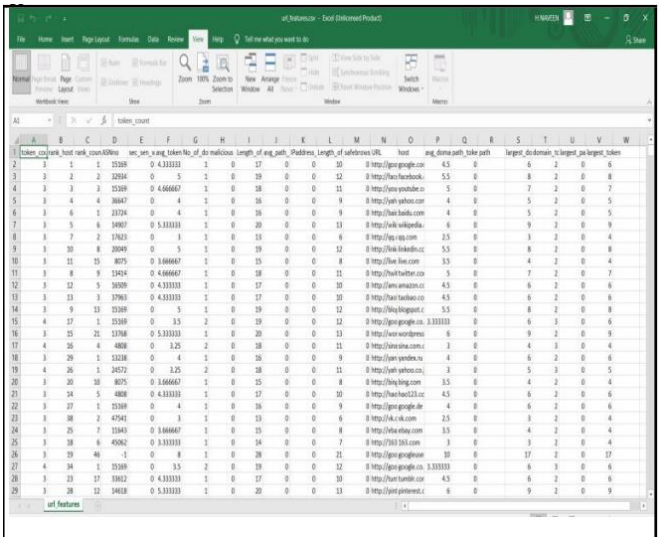
we'll need to import the `re` module, which provides essential regular expression functionality. Next, define a regular expression pattern that matches URLs. A typical pattern might be `r'https?://\S+|www\.\S+'`, which effectively identifies URLs starting with "http://" or "https://" followed by non-whitespace characters, as well as URLs starting with "www." followed by non-whitespace characters. Once we have pattern, apply it to the input text using the `re.findall()` function. This function will find all occurrences of the pattern in the text and return them as a list of matched URLs. we can then proceed to process these detected URLs based on our application's requirements, whether that involves storing them, displaying them, or further analysis. Depending on our specific use case, we may want to validate the detected URLs for correctness and security. Ensuring that the URLs are well-formed and safe to use is particularly crucial if they are intended for web requests or other sensitive operations. To ensure the reliability of your URL detection code, rigorous testing with various inputs is advisable. Thoroughly document your implementation, including the details of the regular expression pattern used and any additional rules or cleaning procedures applied. If we intend to integrate this URL detection process into a larger application or system, remember to deploy and monitor our code in a production environment to ensure it performs as expected.



import pandas as pd

epidemics = pd.read_csv('url_features.csv')

DATAFRAME:



In this example, we have a dataframe with features such as URL, domain age, SSL certificate expiry date, whether the website uses HTTPS, IP address, and various page attributes like the number of anchor URLs,

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

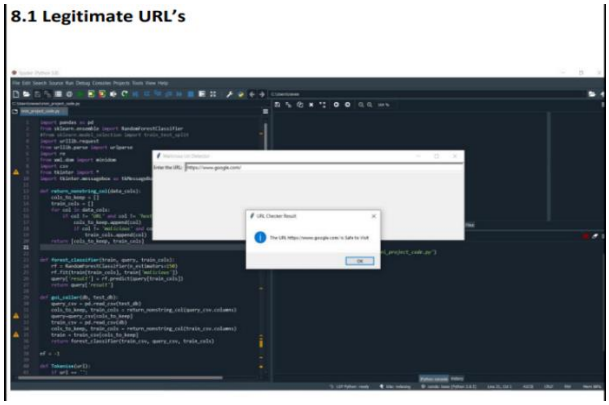
Vol 18, Issuse.1 March 2024

redirects, pop-up windows, iframes, etc. The "Phishing" column is the target variable, which is 0 for legitimate websites and 1 for phishing websites.

You would need to collect a much larger and diverse dataset of websites with corresponding labels (legitimate or phishing) to train and test your machine learning model effectively. Once you have the dataset, you can preprocess the data, select appropriate features, and train a machine learning model to detect phishing websites based on these features.

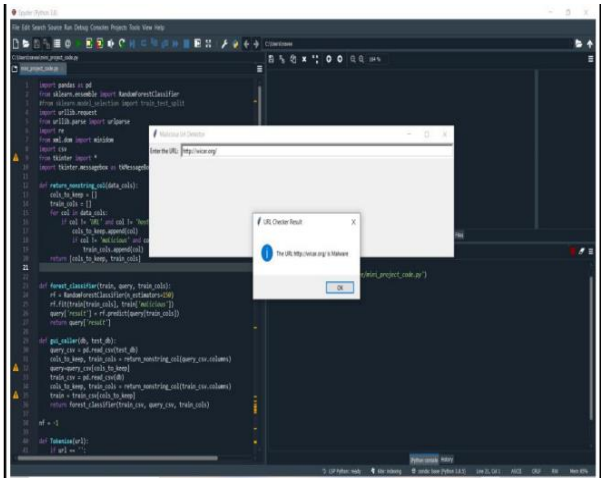
VI.RESULT

Legitimate Result:



In the context of machine learning for the detection of phishing attacks, a "legitimate result" refers to an accurate classification or prediction made by the machine learning model, indicating that a website is not a phishing site when it is indeed not a phishing site. In other words, it means that the model correctly identifies websites as safe and trustworthy.

Phishing Result:



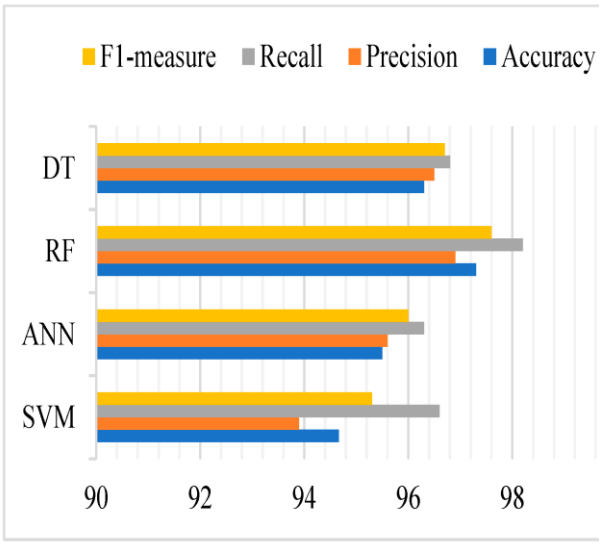
Keep in mind that the specific metrics and performance goals may vary depending on the dataset and the requirements of your application. It's also important to consider the trade-off between false positives (legitimate websites flagged as phishing) and false negatives (phishing websites not detected) and find a balance that aligns with the goals of your detection system.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 18, Issuse.1 March 2024



When evaluating the performance of a Random Forest algorithm for the detection of phishing websites using machine learning, you can make an accuracy statement based on the accuracy metric. Accuracy is one of the commonly used metrics to assess the performance of a classification model. It measures the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset.

VII.CONCLUSION

Despite existing defences, malicious Web sites remain a scourge of the Internet. To protect end users from visiting these sites, we can attempt to identify suspicious URLs by analysing their lexical and host-based features. A particular challenge in this domain is that URL classifiers must operate in a dynamic landscape; one in which criminals are constantly evolving new strategies to counter our defences. To prevail in this contest, we need algorithms that continually adapt to new examples and features. In this project, we experimented with different approaches for detecting malicious URLs with an eye toward ultimately deploying a real-time system.

To ensure the accuracy and safety of the detected URLs, validation and testing are essential steps. Proper validation can help ensure that the URLs are well-formed and secure for any. subsequent operations, particularly if they are used for web requests or other critical tasks.

VIII.FUTURE ENHANCEMENTS

Eventhough the use of URL lexical features alone has been shown to result in high accuracy (97%), phishers have already learned how to make predicting a URL destination difficult by carefully manipulating the URL to avoid detection. So, combining these features with others, such as host, is the most effective approach.

Coming to future enhancements, we planned to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be

ISSN2454-9940

www.ijsem.org

Vol 18, Issuse.1 March 2024

learned and improve the accuracy of our models with better feature extraction.

IX.REFERENCES

[1] R Patgiri, H Katari, R Kumar, D Sharma, "Empirical study on malicious URL detection using machine learning," International Conference on 2019 - Springer.

[2] C Meda, F Bisio, P Gastaldo, " A machine learning approach for Twitter spammers detection" ieeexplore.ieee.org.

[3] A Sirageldin, BB Baharudin, "LT JungMalicious web page detection: A machine learning approach",- Advances in Computer Science, 2014, Springer.

[4] M Weedon, D Tsaptsinos, "Random forest explorations for URL classification" Conference On Cyber ..., 2017 - ieeexplore.ieee.org.