Detection of Phishing Sites

Nageshwar Kauthale, Anushka Ingle, Dhanshri Salve, Madhusudan Vetal

Prof.S.R.Dhotre

---***---

ABSTRACT :

Phishing poses a significant threat to cybersecurity as attackers can easily replicate legitimate websites, tricking users into providing sensitive information. Despite security measures, many users fall victim to these attacks, which result in billions of dollars in losses annually. Traditionally, phishing detection relies on the blacklist method, where known malicious URLs and IPs are stored in databases. However, attackers circumvent these blacklists using techniques like URL obfuscation, fast-flux hosting, and generating new URLs algorithmically. Recent advancements in machine learning offer a promising alternative for detecting phishing websites. By analyzing website content and detecting suspicious patterns, machine learning models trained on large datasets of both legitimate and phishing sites can enhance detection accuracy. Complementary strategies include educating users about recognizing phishing attempts, such as checking URLs for irregularities and avoiding suspicious emails.

KEYWORDS:

Cybersecurity; Social Engineering; Spear Phishing; SVM; Decision Tree; Random Forest.

1. INTRODUCTION

Phishing attacks are a rapidly expanding threat in the cyber world, costing internet users billions of dollars each year. It is a criminal crime that involves the use of a variety of social engineering tactics to obtain sensitive information from users. Phishing techniques can be detected using a variety of types of communication, including email, instant chats, pop-up messages, and web pages. This study develops and creates a model that can predict whether a URL link is legitimate or phishing.

The data set used for the classification was sourced from an open source service called ‘Phish Tank’ which contain phishing URLs in multiple formats such as CSV, JSON, etc. and also from the University of New Brunswick dataset bank which has a collection of benign, spam, phishing, malware & defacement URLs. Over six (6) machine learning models and deep neural network algorithms all together are used to detect phishing URLs. This study aims to develop a web application software that detects phishing URLs from the collection of over 5,000 URLs which are randomly picked respectively and are fragmented into 80,000 training samples & 20,000 testing samples, which are equally divided between phishing and legitimate URLs.

The URL dataset is trained and tested base on some feature selection such as address bar-based features, domain-based features, and HTML & JavaScript-based features to identify legitimate and phishing URLs.

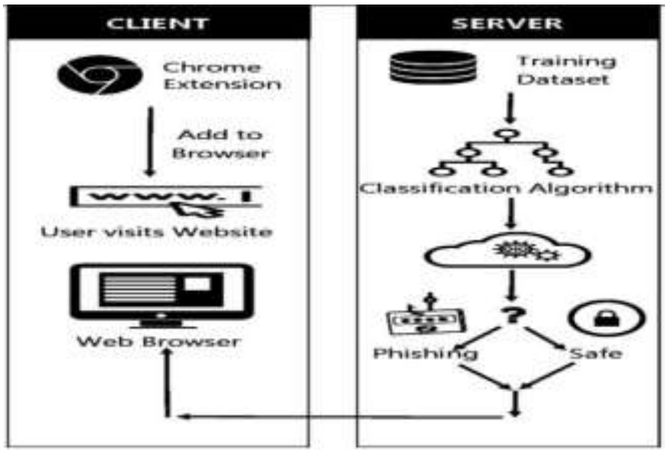
2. NEED:

Phishing site detection is essential in today's digital world due to the growing number of cyber threats targeting individuals and organizations. Phishing websites are designed to mimic legitimate sites in order to deceive users and steal sensitive information such as usernames, passwords, credit card details, and other personal data. Detecting these malicious sites helps protect users from identity theft and financial loss, which can occur through unauthorized transactions or fraudulent activities. For businesses, phishing attacks can damage their reputation, erode customer trust, and lead to legal consequences due to the mishandling of user data. Additionally, phishing sites often serve as platforms for distributing malware, which can compromise systems and lead to broader security breaches. By identifying and blocking phishing sites, organizations can improve their overall cybersecurity posture, ensure compliance with data protection regulations, and contribute to a safer online environment for everyone.

SCOPE OF THE SYSTEM:

The proposed system called ―Detection of phishing sites using Machine Learning‖ enables user to identify particular website is phished website or nonphished website and print the message website is phished or legitimate. We might add the feature of alerting the user about a phished URL via email messages. Text based similarity approaches are relatively fast, but they are unable to detect phishing attack if the text is replaced with some image. Then we might add the feature of detection of the phished images. In the later stage we can implement this approach for android systems..

PROPOSED ARCHITECTURE:



ALGORITHM:

1. Logistic Regression:

Logistic regression is a machine learning algorithm used for binary classification tasks. It estimates the probability that a given input belongs to one of two categories (e.g., yes/no, true/false). The algorithm uses a logistic function to map predicted values to a range between 0 and 1, helping to classify the data.

2. Random forest algorithm:

A set learning regression and category approach called random forest (RF) may be used to solve issues with data classification. Decision trees are used in RF to make predictions. A few decision trees are constructed during the training phase (specified by the programmer) and then utilized for class prediction; this is done by taking into account the graded classes of all individual **trees,** with the class with the highest grade being the output.

3. SVM (Support Vector Machine):

This method is used in the medical industry for illness detection, text content recognition, picture classification, and other purposes. With the use of a fixed rule, a quadratic equation, and statistics, the data will be divided into classes in this way. For the binary classification of the data, a separating hyperplane is utilized, which reduces the space of the margin based on kernel characteristics. This method is employed to identify the ideal response to the issue. This method does not analyze large amounts of data.

IMPLEMENTATION:

- Gathers URLs, website content, and network traffic data.
- Sources: Public datasets (e.g., PhishTank, OpenPhish), web scraping, and user reports.

Join the fight against phishing

Submit suspected phishes. Track the status of your submissions. Verify other users' submissions. Develop software with our free API.

Recent Submissions

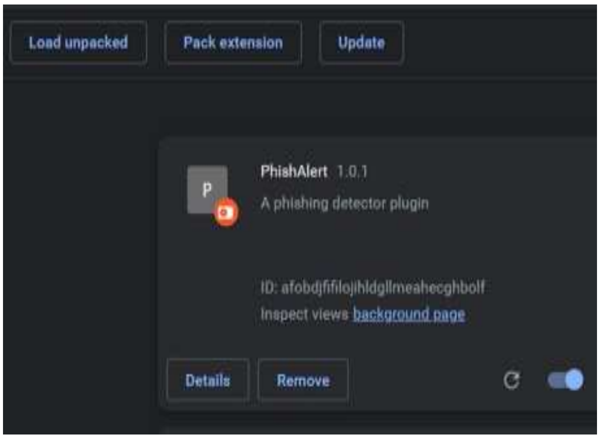
What is phishing?

What is PhishTank?

PhishTank News

Choose Platform

- Develop a **browser extension** for Chrome, Firefox, or Edge.



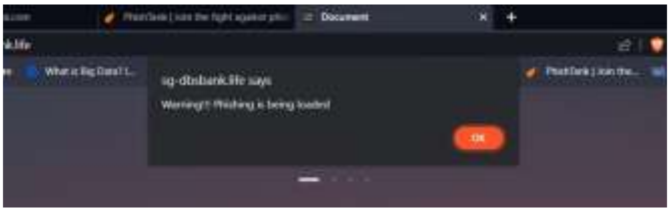
PhishAlert 1.0.1

A phishing detector plugin

ID: afobdjfifilojihldgllmeahecghbolf

Inspect views background page

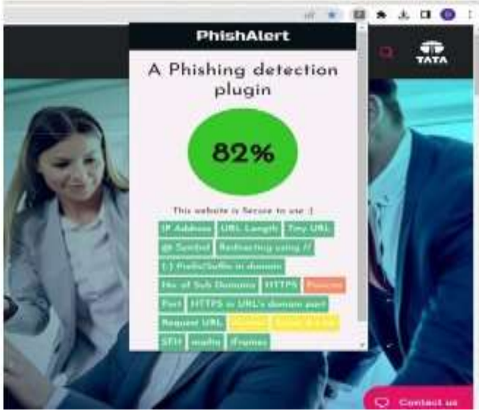
Warning Popup: The output of the plugin while visiting a phishing site taken from Phish Tank. This site has a dialogue box shows warning that you are being phished.



sg-dtnbank.life says

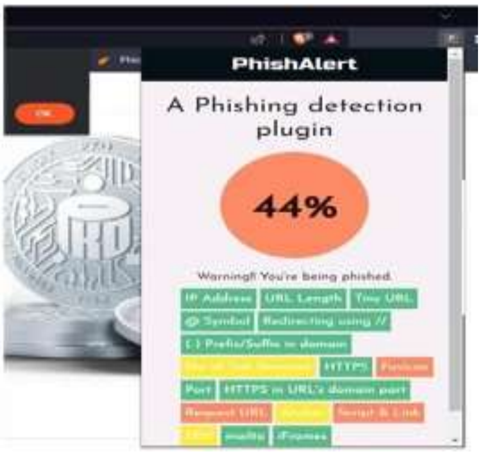
Warning!! Phishing is being loaded

Green Signal: The output of the plugin while visiting a phishing site taken from Phish Tank. This site has a high trust value and also the light red circle indicates phishing or not.



This website is Secure to use :)

Red Signal: The output of the plugin while visiting a phishing site taken from Phish Tank. This site has a low trust value and also the light red circle indicates phishing



Warning!! You're being phished.

METHODOLOGY:

Data Collection: Phishing and legitimate website data are collected from various sources, such as phishing databases or user reports.

Preprocessing: The collected data undergoes preprocessing, which may include cleaning, filtering, and transforming the data into a suitable format for further analysis. Feature

Extraction: Relevant features are extracted from the website data, such as URL properties, HTML content, metadata, or behavioral patterns.

Training Data Preparation: The extracted features are combined with labels indicating phishing or legitimate status to create a labelled training dataset. Machine Learning Model

Training: The training dataset is used to train a machine learning model, such as SVM, random forest, or decision trees. The model learns the patterns and characteristics associated with phishing and legitimate websites. Model

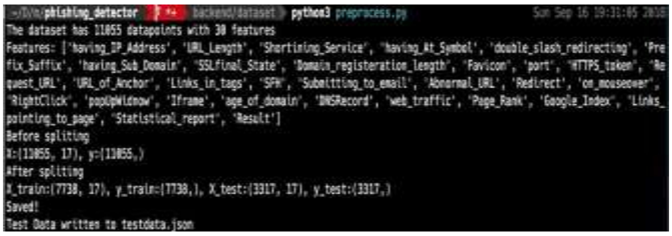
Reporting and Analysis: The system can generate reports and provide insights on detected phishing attempts, false positives, and other relevant statistics for further analysis and improvement.

DEPLOYMENT

In the deployment phase, the application is made available to users. Many companies prefer to automate the deployment phase. This can be as simple as a payment portal and download link on the company website. It could also be downloading an application on a smartphone. NBNSTIC, Computer Science Engineering 2022-23 30 Deployment can also be complex. Upgrading a companywide database to a newly-developed application is one example. Because there are several other systems used by the database, integrating the upgrade can take more time and effort.

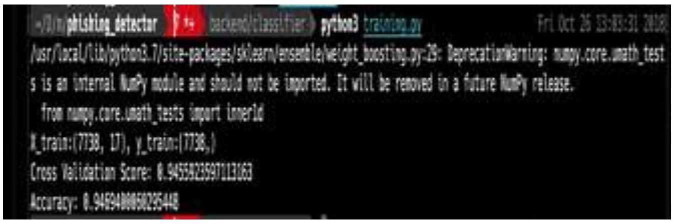
BACKEND PROGRAM:

PREPROCESSING



```
~/D/m/phishing_detector  backend/dataset  python3 preprocess.py                Sun Sep 16 19:31:05 2018
The dataset has 11055 datapoints with 30 features
Features: ['having_IP_Address', 'URL_Length', 'Shortining_Service', 'having_At_Symbol', 'double_slash_redirecting', 'Pre
fix_Suffix', 'having_Sub_Domain', 'SSLfinal_State', 'Domain_registeration_length', 'Favicon', 'port', 'HTTPS_token', 'Re
quest_URL', 'URL_of_Anchor', 'Links_in_tags', 'SFH', 'Submitting_to_email', 'Abnormal_URL', 'Redirect', 'on_mouseover',
'RightClick', 'popUpWidnow', 'Iframe', 'age_of_domain', 'DNSRecord', 'web_traffic', 'Page_Rank', 'Google_Index', 'Links_
pointing_to_page', 'Statistical_report', 'Result']
Before splitting
X:(11055, 17), y:(11055,)
After splitting
X_train:(7738, 17), y_train:(7738,), X_test:(3317, 17), y_test:(3317,)
Saved!
Test Data written to testdata.json
```

TRAINING



```
~/D/m/phishing_detector  backend/classifier  python3 training.py              Fri Oct 26 13:03:31 2018
/usr/local/lib/python3.7/site-packages/sklearn/ensemble/weight_boosting.py:29: DeprecationWarning: numpy.core.umath_test
s is an internal NumPy module and should not be imported. It will be removed in a future NumPy release.
  from numpy.core.umath_tests import inner1d
X_train:(7738, 17), y_train:(7738,)
Cross Validation Score: 0.9455823597113163
Accuracy: 0.9459400060295448
```

The project underwent peer review to ensure that the design and implementation were sound. Feedback was gathered from several peers and experts in the field, embedded systems, and safety systems

FUTURE SCOPE:

The future of phishing site detection lies in the use of advanced technologies like AI and machine learning for faster and more accurate real-time detection. These tools will help identify phishing websites by analyzing patterns, behavior, and design. Integration with browsers and email platforms will provide users with instant warnings. Technologies like blockchain may also be used to verify website authenticity, while global threat sharing and increased user awareness will play an important role in improving overall cybersecurity.

CONCLUSIONS

In conclusion, phishing continues to evolve as cybercriminals adopt more sophisticated techniques to deceive individuals and organizations. The integration of AI, deepfakes, and automation into phishing attacks presents significant challenges, while the increasing reliance on cloud services, IoT devices, and cryptocurrency platforms expands the potential attack surface. As technology progresses, phishing will likely target new areas such as mobile devices, blockchain systems, and 5G networks. However, advancements in AI-based detection, biometric authentication, and quantum-resistant encryption offer promising avenues for defense. Addressing both the technological and human factors is essential for developing effective countermeasures against the future of phishing.

REFERENCES:

1. **"Phishing Detection: A Literature Survey"**
   Authors: Mahmoud Khonji, Youssef Iraqi, Andrew Jones
   Summary: This survey provides a comprehensive overview of various phishing detection techniques, analyzing their effectiveness and limitations. *Link:* Phishing Detection: A Literature Survey
2. **"Life-long phishing attack detection using continual learning"**
   Authors: [Authors not specified in the provided snippet]
   Summary: This paper explores continual learning techniques to maintain phishing detection performance over time, addressing the evolving nature of phishing attacks. *Link:* Life-long phishing attack detection using continual learning
3. **"An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs"**
   Authors: Ayman El Aassal, Shahryar Baki, Avisha Das, Rakesh M. Verma
   Summary: This study conducts a systematic benchmarking of phishing detection features and methods, providing insights into their performance across diverse datasets. *Link:* An In-Depth Benchmarking and Evaluation of Phishing Detection Research
4. **"Unveiling suspicious phishing attacks: enhancing detection with an optimal feature vectorization algorithm and supervised machine learning"**
   Authors: Maruf A. Tamal, Md K. Islam, Touhid Bhuiyan, Abdus Sattar, Nayem Uddin Prince
   Summary: This study aims to develop a robust solution for phishing detection through optimal feature vectorization and supervised machine learning classifiers. *Link:* Unveiling suspicious phishing attacks
5. **"Phishing URL detection with neural networks: an empirical study"**
   Authors: [Authors not specified in the provided snippet]
   Summary: This paper reports on the application of deterministic and probabilistic neural network models to URL classification for phishing detection. *Link:* Phishing URL detection with neural networks
6. **"A comprehensive survey of AI-enabled phishing attacks detection techniques"**
   Authors: [Authors not specified in the provided snippet]
   Summary: This paper provides a literature review of Artificial Intelligence techniques, including Machine Learning, Deep Learning, and Hybrid Learning, applied to phishing detection.

Link: A comprehensive survey of AI-enabled phishing attacks detection techniques

7. **"A comprehensive literature review on phishing URL detection using deep learning approaches"** *Authors:* [Authors not specified in the provided snippet] *Summary:* This analysis examines the state-of-the-art in deep learning for phishing URL identification, reviewing 41 research studies from 2019 to 2024. *Link:* A comprehensive literature review on phishing URL detection using deep learning approaches
8. **"Phishing Website Detection through Multi-Model Analysis of HTML Content"** *Authors:* Furkan Çolhak, Mert İlhan Ecevit, Bilal Emir Uçar, Reiner Creutzburg, Hasan Dağ *Summary:* This study introduces an advanced detection model focusing on HTML content, integrating multiple models for improved hishing website detection. *Link:* Phishing Website Detection through Multi-Model Analysis of HTML Content
9. **"Towards Web Phishing Detection Limitations and Mitigation"** *Authors:* Alsharif Abuadbba, Shuo Wang, Mahathir Almashor, Muhammed Ejaz Ahmed, Raj Gaire, Seyit Camtepe, Surya Nepal *Summary:* This paper explores the limitations of current phishing detection methods and proposes a more resilient model incorporating additional feature spaces. *Link:* Towards Web Phishing Detection Limitations and Mitigation
10. **"A Sophisticated Framework for the Accurate Detection of Phishing Websites"** *Authors:* Asif Newaz, Farhan Shahriyar Haq, Nadim Ahmed *Summary:* This paper proposes a comprehensive methodology for detecting phishing websites, utilizing feature selection, greedy algorithms, and deep learning methods. *Link:* A Sophisticated Framework for the Accurate Detection of Phishing Websites