AN EFFECTIVE PHISHING WEBSITE DETECTION USING MACHINE LEARNING

Sakeena khan, Mtech student of CSE,Shadan Women's College of Engineering, Hyderabad,telangana,India. ksakeena068@gmail.com.

Dr. K.Palani,Professor and Principal, Department of CSE, Shadan Women's College of Engineering, Hyderabad,telangana,India.principalswcet2020@gmail.com.

Dr.P. Senthil Kumar,Professor, Department of CSE, Shadan Women's College of Engineering, Hyderabad,telangana,India.psenthilkumarshadan@gmail.com.

ABTRACT

Websites with malicious intent slow down the development of Web services and greatly promote the growth of online crime. So, there's been a lot of effort to come up with methods to systematically prevent people from going to those kind of websites. We propose a learning-based system for classifying webpages as either benign, spam, or harmful. Instead of accessing actual website content, our method only looks at the URL. Consequently, it eliminates the possibility of run-time latency and browser-based vulnerabilities. Using learning techniques, our approach achieves better coverage and generalizability than blacklisting services.

A website may use one of three distinct types of URLs:

- **Non-threatening: safe websites providing basic services**
- **Spam: Some websites aim to flood users with ads or other information, including fake surveys or online dating, in an effort to deceive them.**
- **Websites created by hackers with the goal of stealing private information, disrupting computer operations, or both are known as malware.**

INTRODUCTION

While Because of the Internet, it is now simpler than ever before for many people to keep track of their money and possessions, it has also made it easier than ever for large-

scale, low-cost fraud to occur. It is increasingly possible for fraudsters to manipulate individuals rather than software or hardware systems due to the decreased barriers to technological entry. In the realm of internet fraud, phishing is among the most common forms. Credit card numbers and passwords are among the sensitive data that are targeted for theft. A phishing attack may take two forms.

Endeavours to infect victims' systems with malware for the purpose of stealing secrets by pretending to be trustworthy businesses that really want the information.

Since the specific malware that is used in phishing scams includes an area of research for the virus and malware community, it is not included in this thesis. This thesis delves at the study around deceptive email marketing tactics, often known as a "phishing attack."

TYPES OF PHISHING

Deceptive Phishing: This is the most common kind of malicious email attacks, where the criminal poses as a legitimate business or organization with the hope of fooling the victim into divulging sensitive information. Because it does not need customization or modification, this kind of assault is simpler.

Spear Phishing: Along with sensitive information on the potential victim, this kind of phishing email includes malicious URLs. Details about the recipient's social media accounts, friends, colleagues, and job title might end up in the inbox.

Whale Phishing: In an effort to spear phish a "whale," or high-ranking executive such as a chief executive officer or other top executive, this kind of phishing aims to target such individuals. In URL phishing, the scammer or cybercriminal infects the victim by means of a URL link. People are so ready to hit the "accept friend" button because of their innate extroversion. is ready to provide private details like email addresses and perhaps even issue an invitation. This is the result of phishing attempts that lead victims to a bogus server. Criminals also employ encrypted browser connections to do their illegal acts. Attacks like this become more common because companies can't afford to teach their staff about phishing, which makes it tougher for them to stop them.Companies are being cautious by encrypting sensitive data, keeping all systems updated with the newest security patches, and teaching employees via simulated phishing attacks.

One certain method to fall for this phishing scam is to be careless while perusing the web. Phishing websites seem very much like legitimate ones.

OBJECTIVE

Using deep learning to identify dangerous, malicious, and start URLs is the primary objective of this work.

MOTIVATION

To safeguard users from these harmful websites, this procedure was put into place. Public awareness will be increased, and strong security measures will be put in place to detect and prevent phishing URLs from reaching users.

LITERATURE SURVEY

TITLE-1: Automated Classification of Phishing Pages on a Large Scale.

AUTHOR: Colin Whittaker, Brian Ryner, Marria Nazif.

YEAR : 2010

The yearly cost to Internet consumers due to phishing websites exceeds one billion dollars. Websites that engage in phishing attempt to trick users into divulging critical information by pretending to be legitimate third parties. A scalable ML classifier that can identify fraudulent websites is presented in this paper, along with its architecture and functionality. With the help of this classifier, Google's phishing blacklist will be automatically updated. Our classifier checks the URL and content of millions online publications daily to determine whether they are phishing. Unlike earlier studies in this field, the classifier was trained using million-sample dataset with noise from actual categorization data that had already been collected. In spite of the fact that the training data contains noise, our classifier is able to construct a robust model for phishing page detection; weeks after training finishes, it still manages to correctly categorize more than 90% of phishing sites.

TITLE-2: Using an Intrusion Detect Dataset for Machine Learning Algorithms in a Misuse Detection Setting.

AUTHOR:MaheshkumarSabhnani, Gursel Serpen.

YEAR: 2003

According to the present research, the KDD 1999 Cup intrusion detection dataset produced terrifying findings for the distinct types of attacks, such as user-to-root and

remote-to-local, where an small subset of machine learning algorithms, mostly focused on inductive learning, were used. The work described here is motivated by the ambiguity around the possible superiority of alternative machine learning algorithms over existing ones. It is of particular interest to investigate if there are any algorithms that outperform others for certain types of attacks, and if so, whether this might lead to the development of a multi-expert classifier that satisfies the required performance requirement. Using four distinct attack types from the KDD 1999 Cup intrusion detection dataset, this research assesses the efficacy of several machine learning and pattern recognition methods. In this setting, a simulation research shown that some classification algorithms perform better against certain types of threats.

TITLE-3: Making Efficient Classifiers to Identify Spam Images.

AUTHOR: Mark Dredze, Reuven Gevaryahu, Ari Elias-Bachrach.

YEAR: 2007

Rather of having the spam text shown in the message body, spammers have begun sending out more "image spam" recently. This kind of email makes it tougher for typical content filters to detect the spam. We need innovative approaches in order to screen out signals. The goal is to classify images as spam or ham in an instant and automatically. To facilitate speedy picture classification by showcasing features that emphasize their essential qualities. The test results show that spam images can be correctly categorized more than 90% of the time when utilizing real-world data, and up to 99% of the time when using artificial data. Moreover, a new feature selection method that prioritizes fast and accurate prediction while selecting features for classification. This approach yields a precise system that runs in a negligible amount of time. Justin Time (JIT) feature extraction is included at the end to showcase just-in-time extraction with a JIT decision that improves system performance even more. JIT feature extraction creates features when the classifier needs them. To make photo spam classification a reality, this work offers a fast learning technique for classifiers and features with excellent accuracy.

EXISTING SYSTEM

Without proper structure in the NN model, the training dataset runs the risk of being underfit. However, overfitting might occur if you attempt to have the algorithm

match every object in the training dataset. Changing certain values, enhancing the buried layer with more neurons, or introducing a network layer periodically is one way to restructure the NN model and prevent the Overfitting issue. A NN with a limited amount of hidden neurons could struggle to capture the variety and depth of the data. Conversely, networks that include an excessive number of hidden neurons are more likely to experience overfitting. The model can only be improved to a certain point before the structural process must terminate. Because it's hard to predict what an acceptable error rate is, a satisfactory rate of mistake must be supplied when developing that uses NN. For example, if the designer sets the allowed error rate to an unrealistically high or unrealistically low value, the model can get stuck in a local minimum.

DISADVANTAGES

- Loading the whole dataset will take some time.
- The procedure is inaccurate.
- The analysis will be gradual.

PROPOSED SYSTEM

The fact that legitimate and illegal websites often use distinct URLs is the source of lexical characteristics. Analyzing lexical characteristics allows us to capture the property for the purpose of classification. Get the route and host name out of a URL before you can extract the bag-of-words. Find out that phishing websites like longer tokens, more levels, more tokens in the domain and route, and longer URLs. In addition, malicious and phishing websites may masquerade as legitimate ones by employing popular brand names as tokens that are distinct from those in second-level domains. Considering that normal websites almost never use IP addresses to conceal suspicious URLs, phishing and malicious websites often do so. Also, a lot of suggestive word tokens are seen in phishing URLs.In order to check for the presence of certain security-sensitive phrases and add the binary value to our features. By definition, fewer people visit malicious websites than safe ones. One may thus argue that website popularity is crucial. An analysis of Alexa traffic rank was conducted. A host-based feature is based on the idea that malicious websites are often registered in less trustworthy hosting places.

ADVANTAGES

- Labels are attached to each URL in the collection.

- For training, two Random forests and support vector machines are two examples of supervised learning techniques.—are used inside the scikit-learn package.

METHODOLGY OF PROJECT

Four approaches are being used in this case.

- Collecting Data
- Data Preparation
- Notable characteristics
- Assessment Model

ALGORITHMS USED

RANDOM FOREST

When faced with a regression or classification issue, one well-known supervised machine learning approach is random forest. In order to construct decision trees, several samples are used. For classification, the majority vote is used, while for regression, the average vote is used.The ability of the Method of Random Forests for Dealing with Data Sets that include both continuous and categorical variables is one of its most notable characteristics. Continuous variables are the ones used in regression, whereas classification variables are more often seen. It works better when classification is involved.We may look at a real-life example to help us understand this better. X, the student, is uncertain about the degree to follow after finishing his 10+2 since he does not know which skills to put to use. Therefore, he decides to talk to a wide range of people, including his parents, teachers, relatives, other students, and adults in the workforce. Course expenses, career prospects, and his decision-making process are just a few of the subjects he probes them with questions on. After consulting with many people, he decides to sign up for the class that came highly recommended.

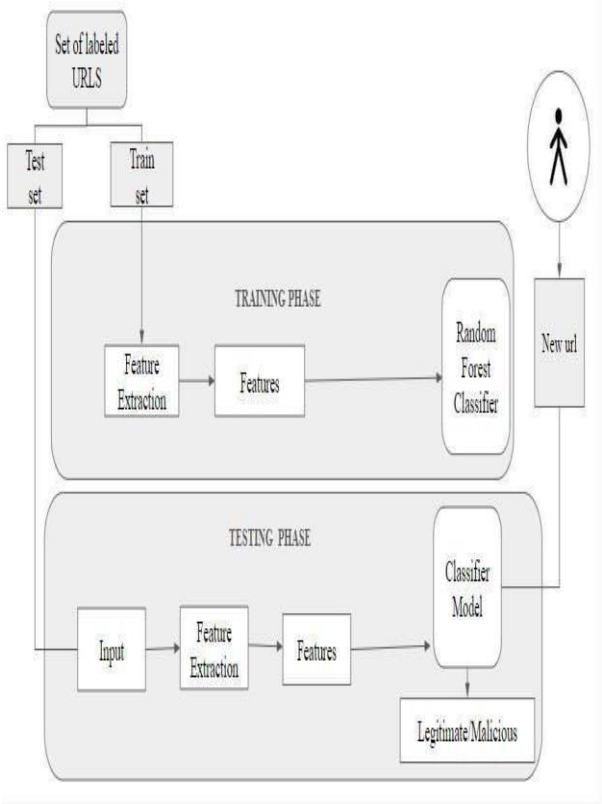
SUPPORT VECTOR MACHINE

For jobs like regression, finding outliers, and linear or nonlinear classification, strong ML techniques are used, such as The SVM is a support vector machine. Automated support vector machines have several uses, in areas such as text categorization, picture categorization, handwriting identification, spam identification, facial identification, gene expression analysis, unusual event detection, and many more. The adaptability and efficacy of support vector machines (SVMs) stem from their ability to process data with high dimensions and nonlinear relationships.

DECISION TREE

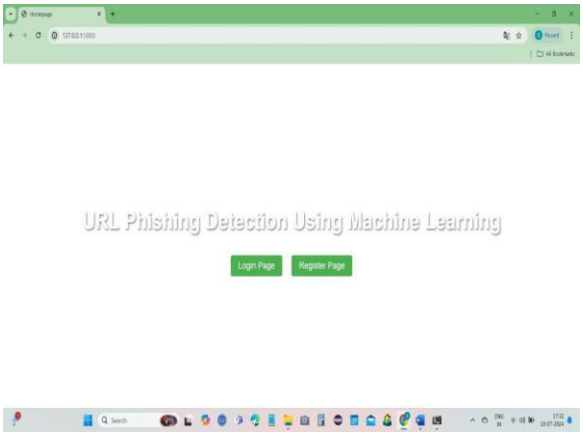
One adaptable and easily understood method for predictive modeling in machine learning is a decision tree. Its judgments are organized according to input data, making it suitable for both classification and regression tasks. In addition to discussing decision trees and their learning algorithms, this page breaks them down into their component elements, terminology, construction, and advantages.

SYSTEM ARCHITECTURE

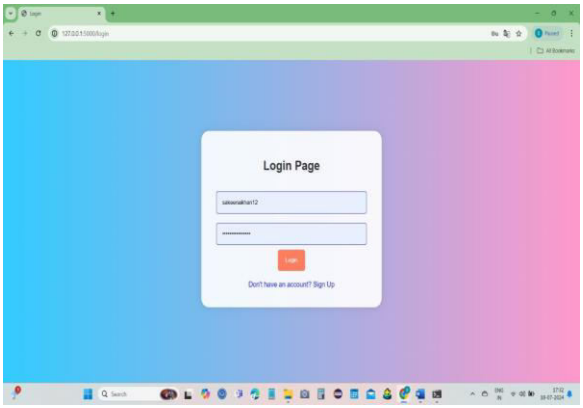


SYSTEM ARCHITECTURE

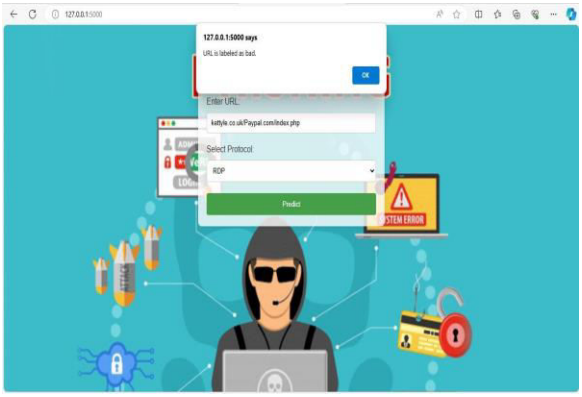
EXPERIMENTAL RESULTS



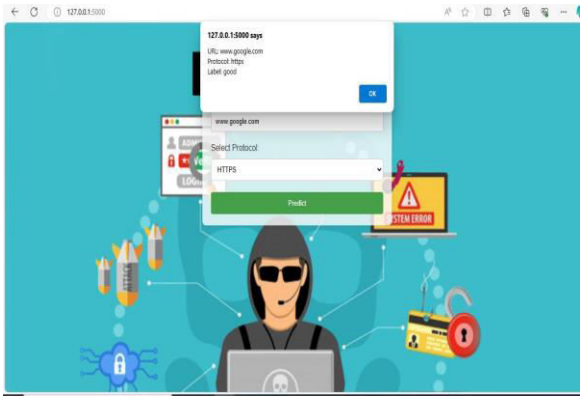
HOME PAGE



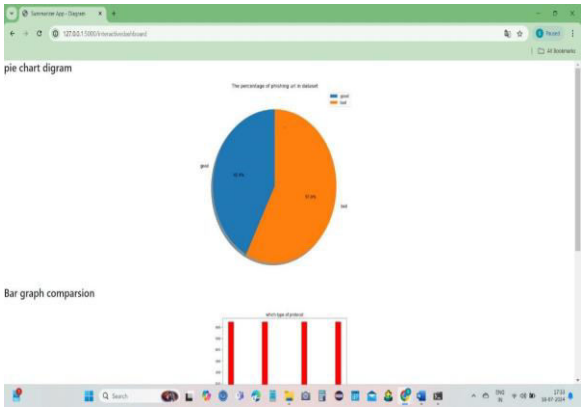
LOGIN PAGE



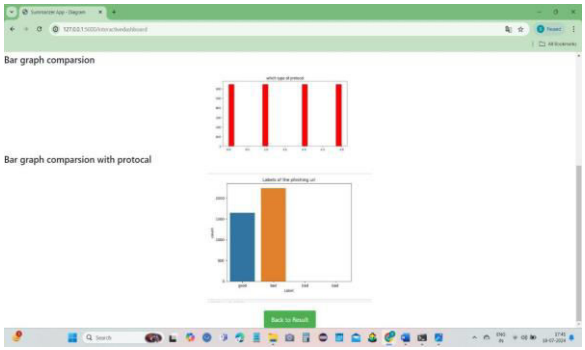
DETECTION OF URL IS BAD



DETECTION URL IS GOOD



PIE CHART DIAGRAM



BAR GRAPH COMPARISON

CONCLUSION

To the tune of zero false positives.1%, our large-scale technology for automatically identifying phishing websites is described in this project. A number of promising areas to guide investigations into the field of machine learning-based phishing URL detection include:

Developing more sophisticated models:

To combat more sophisticated phishing URLs—including those that use obfuscation techniques—researchers may train more advanced machine learning models.

Incorporating more features:

To enhance the precision of phishing URL identification, researchers may augment their models with details like study of online material, examination of data flow over networks, and study of user actions.

FUTURE WORK

Numerous methods exist for a machine learning-based phishing website detection project might be improved. A few characteristics might added or changed with new ones to identify phishing websites, which are becoming more common by the day.

When working on a machine learning project to improve phishing website detection, it's important to look at several advanced approaches and tactics that may make the project more efficient, accurate, and flexible. A few possible upgrades for the future are these:

Advanced Feature Engineering

• **Use of Natural Language Processing (NLP):** Keep an eye out for deceptive methods and misleading wording on phishing websites.

Hybrid Models

- **Ensemble Learning:** Improve detection accuracy by combining numerous models and using the capabilities of diverse algorithms.
- **Transfer Learning:** Reduce the requirement for big datasets by using pre-trained models on comparable tasks and adapting them to phishing detection.

Real-Time Detection and Response

- **Streaming Data Analysis:** Put in place systems that can evaluate data from websites to identify phishing attempts as they happen.
- **Adaptive Learning:** Create models that can learn new phishing tactics and strategies from incoming data and adjust their strategies and tactics accordingly.

Integration with Web Browsers and Security Tools

- **Browser Extensions:** Make add-ons for web browsers that use machine learning algorithms to detect and block phishing websites.
- **API Integration:** For better defense, combine detection systems with preexisting security solutions and databases.

User Education and Feedback Mechanisms

- **User Feedback Integration:** Create systems that allow the detection system to include user comments and input such that a enhance the performance of the model on an ongoing basis.

REFERENCES

- B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, “Fighting against phishing attacks: state of the art and future challenges,” Neural Computing and Applications, vol. 28, no. 12, pp. 3629–3654, 2017.
- Lakshmanarao, A., Rao, P.S.P., Krishna, M.M.B. (2021) ‘Phishing website detection using novel machine learning fusion approach’, in 2021 International Conference on Artificial Intelligence and

Smart Systems (ICAIS), Presented at the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 1164– 1169.

- Sci-kit learn, SVM library. http://scikit-learn.org/stable/modules/svm.html.
- survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013. G. Park and J. M. Taylor, "Using syntactic features for phishing.
- Vayansky, I. and Kumar, S., "Phishing – challenges and solutions.", Computer Fraud & Security, vol 2018, no. 1, pp. 15-20, January 2018.
- Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, "Phishing Detection Using Machine Learning Techniques," unpublished.
- Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. "Machine Learning-Based Phishing Detection from URLs," Expert Systems with Applications, vol. 117, pp. 345357, January 2019.
- J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," International Conference on Control Communication and Computing (ICCC), December 2013. Pradeepthi, K. V., & Kannan, A. "Performance study of classification techniques for phishing URL detection," Sixth International Conference on Advanced Computing (IcoAC), December 2014.