Phishing Detection System using Random Forest Algorithm

1Dr.G. Ramesh, 2R.B. Lokitha, 3R.R. Monisha, 4N.S. Neha

1Professor, Department of Information technology, K L N College of Engineering, Sivagangai, Tamil Nadu, India.
2,3,4Student, Department of Information technology, K L N College of Engineering, Sivagangai, Tamil Nadu, India.

Abstract— The Internet has integrated easily into our daily lives, but it has also made it possible to carry out illicit activities like hacking in secret. Phishers attempt to trick their victims by employing social engineering techniques or building fake websites to take information such as account IDs, usernames, and passwords from people and businesses. Phishing is a form of electronic online identity theft in which the attackers deceive a user into divulging sensitive information by using a mix of social engineering and website spoofing techniques. Typically, this knowledge is employed to generate an unlawful financial gain. (e.g., by online banking transactions, purchase of goods using stolen credentials, etc.). A remarkable platform for communication between regular individuals is the Internet. People with criminal intent have discovered a method to steal people’s personal information with the least chance of being caught and without meeting the target. It’s known as scammers. The e-commerce sector is seriously threatened by phishing. Customers’ trust in online shopping is undermined, and electronic service suppliers suffer significant financial losses as a result. For this reason, understanding malware is crucial. Although many strategies have been put forth to identify scam websites, phishers have developed ways to circumvent these strategies. Machine learning is one of the best techniques for spotting these malevolent behaviors. This is because most phishing attacks share some similar traits that can be.

Keywords—Phishing, Classification, Optimization, Map Reduce.

INTRODUCTION

Online services simplify our lives. They allow us to access information ubiquitously and are also useful for service providers because they reduce the operational costs involved in offering a service. For example, online banking over the web has become indispensable for customers as well as for banks. Unfortunately, interacting with an online service such as a banking web application often requires a certain degree of technical sophistication that not all Internet users possess. For the last couple of years, such naive users have been increasingly targeted by phishing attacks that are launched by miscreants who are aiming to make an easy profit by means of illegal financial transactions. Phishing is a form of electronic identity theft in which a combination of social engineering and web site spoofing techniques are used to trick a user into revealing confidential information with economic value. In a typical attack, the attacker sends a large number of spoofed (i.e., fake) emails to random Internet users that appear to be coming from a legitimate business organization such as a bank. The e-mail urges the recipient (i.e., the potential victim) to update his personal information. Often, the e-mail also warns the recipient that the failure to comply with the request will result in the suspension of his online banking account. Such ungrounded threats are common in social engineering attacks and are an effective technique in persuading users. When the unsuspecting victim follows the phishing link that is provided in the email, he is directed to a web site that is under the control of the attacker. The site is prepared in a way such that it looks familiar to the victim. That is, the phishers typically imitate the visual corporate identity of the target organization by using similar colors, icons, logos and textual descriptions. In order to “update” his personal information, the victim is asked to enter his online banking login credentials (i.e., username and password) to access the web site. If a victim enters his valid login credentials into the fraudulent web site, the phisher can then impersonate the victim. This may allow the attacker to transfer funds from the victim’s account or cause other damage. Because victims are directly interacting with a website that they believe they know and trust, the success rate of such attacks is very high. Note that although phishing has been receiving wide media coverage (hence, causing the number of Internet users who have heard of phishing to increase), such attacks still remain effective as phishers have been adapting their social engineering attempts accordingly. For example, many phishing emails now ask the victims to validate their personal information for “security purposes”, supposedly because the targeted organization would like to protect them against the phishing threat.

1. METHODOLOGY

Random forest is a machine learning algorithm for fraud detection. It’s an unsupervised learning algorithm that identifies fraud by isolating outliers in the data. This random partitioning of features will produce shorter paths in trees for the fraud data points, thus distinguishing them from the rest of the data. In this system, the user will login using his/her user ID and password. After successful authentication, the user will be directed to the home page where the user can enter the suspicious URL which needs to be tested. After entering the URL, the first step is to check whether the database has already been an entered URL which is marked as phishing. If found, the available result or information of the corresponding URL is obtained from the database and presented before the user in the GUI of the user. Else, the system takes the URL to be tested by using machine learning technique. It will take the URL and classify using Random Forest technique, which is a popular machine learning technique used for classification processes. After the classification, if the URL turns out to be a phishing site, the resultant information will be published in the GUI.
Dataset collection & data preprocessing

Attribute values are in the form of integer -1, 0 and 1 where -1 represents legitimate, 0 represents doubtful and 1 represents phishing. First, the dataset has been pre-processed to split the data in desired format. Then it is divided into two sections as trained data and test data. The experiment has been carried out using java to detect phishing URL based applications. Random forest algorithm has been used in the work for detection of phishing websites.

Random forest classification

The dataset is split into trained data and test data. The dataset is reduced by 80% to get trained data, and then the trained data is again reduced by 25% to get test data. One of the many uses for the well-known machine learning algorithm Random Forest is the detection of phishing URLs. The algorithm builds a “forest” of decision trees by first building several of them. To help prevent overfitting, each decision tree is trained on a portion of the data and employs a random selection of features.

You need a dataset of URLs that have been classified as phishing or legitimate before using Random Forest to detect phishing URLs. From URLs, you can gather a variety of information, including the length of the URL, the number of subdomains, the presence of particular keywords, and more. The Random Forest model receives these characteristics as input.

Prediction

The steps outlined below have to be followed in order to use a trained Random Forest model to determine whether a new URL is phishing or not:

1. Preprocess the new URL: Extract relevant URL features such as length, number of subdomains, presence of specific keywords, and so on. Using the same preprocessing steps that were used on the training data.
2. Feed the features into the Random Forest model: The trained Random Forest model will therefore receive the new URL’s extracted features as input.
3. Get the predicted label: The output of the Random Forest model will be a prediction of whether the new URL is phishing, doubtful or legitimate.
4. Interpret the predicted label: If the predicted label is 1, then the model has predicted that the new URL is phishing. If the predicted label is 0, then the model has predicted that the new URL is legitimate.

II. IMPLEMENTATION

![Diagram of dataset processing flow](image)
1. Dataset collection

Establish the difference between sources that contain legitimate URLs, such as well-known websites or social media sites, and sources that contain phishing URLs, such as public phishing repositories or security reports. Gather a list of URLs from the selected sources. To gather URLs from websites or from social media platforms, use web scraping techniques or APIs. Indicate the legitimacy or phishing of each URL. The URLs can be automatically labeled or manually labeled using pre-existing labeled datasets. Remove duplicate URLs and invalid URLs from the dataset. Extract a set of characteristics from each URL, such as the URL's length, the presence of specific keywords, and the domain name. Preprocess the data, including scaling or normalizing the features if necessary. Save the preprocessed data in a suitable format, such as CSV file.

2. Data preprocessing

Load the preprocessed dataset in a suitable format, such as CSV file, using a data processing library. Remove duplicate URLs. Check and remove any duplicate URLs from the dataset. Remove invalid URLs too. Split the dataset into training and test dataset. Scale the features in the dataset. Save the preprocessed data in a CSV file.

3. Hadoop splitting

The MapReduce framework in Hadoop is used to divide the dataset into smaller pieces so that multiple nodes can handle them concurrently. Every node executes two functions: a Map function to extract features from the URLs, and a Reduce function to use those features to train a random forest model. The end result is a combined model that can be used to identify phishing.

4. Random Forest Classification

Since it can manage massive data sets and complex feature spaces, the random forest algorithm was preferred by us. The algorithm builds numerous decision trees on various random subsets of the data, then combines the output from each tree. We distributedly implemented the random forest method using Apache Spark, an open-source library. The model was trained using training data, with hyper-parameters tuned using cross-validation. Random Forest isolates fraud in the data points instead of profiling non-fraud data points. As fraud data points mostly have a lot shorter tree paths than the normal data points, trees in the isolation forest do not need to have a large depth so a smaller max_depth can be used resulting in low memory requirement.

5. Performance analysis

It’s a process of Detection of phishing URLs from the dataset. The malicious web pages from the data sets are effectively enhanced with the overall performance. The overall performance of the system is increased leading with the accurate prediction of results.

6. Prediction

The proposed approach outperformed other modern phishing detection techniques and obtained high accuracy. Using the random forest method, we were able to pinpoint key characteristics that serve as key indicators of phishing websites. Our suggested model can be an effective weapon in the struggle against phishing scams and can enhance defense.
III. CONCLUSION

Due to the importance in safeguarding privacy and ensuring security, researchers are now very interested in the detection of phishing. There are numerous techniques for classifying websites using trained machine learning models to identify phishing. The speed of detection is increased by URL-based research. There are numerous machine learning algorithms with excellent performance metrics for classification. The method of phishing detection and the phishing detection schemes in the most current research literature have both been studied in this paper.

This paper aims to enhance detection methods to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using the random forest algorithm with the lowest false positive rate. Also, the result shows that classifiers give better performance when we use more data as training data. In future, hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

IV. REFERENCES


