WEBSHIELD: Phishing Website Detection using Machine Learning

Soumya A[1], Chandrashekar KM[2], Prajwal[3], Vijay Kumar S Biradar[4], D Prashanth Reddy[5]

Department of Information Science and Engineering, Rao Bahadur Y Mahabaleshwarappa Engineering College

---***---

Abstract - Phishing is an online hoax whereby an intruder sends fraudulent messages which appear to be from a credible source. There will be included in the letter a URL or document, clicking which will capture private details or install a virus on a computer. Phishing traditionally used to take place by sending mass-scale spam campaigns targeting large populations randomly. The objective was to make as many individuals click on a link or open an infected file as possible. There are different methods to identify this kind of attack. One of the methods is machine learning. The URLs received by the user will be input to the machine learning model then the algorithm will execute the input and show the output whether it is phishing or not.There are different ML algorithms such as SVM,Neural Networks, Random Forest, Decision Tree, XG boost etc. which can be employed to classify these URLs. By comparing and extracting various features between legitimate and phishing URLs, the proposed method employs gradient boosting classifier to detect phishing URLs. The findings of the studies prove that the proposed approach effectively identifies legitimate websites from fake ones in real time

Key Words: URL classification, Machine Learning, Spam campaigns, Gradient Boosting Classifier, SVM, Neural Networks, Real-time detection, Feature extraction

1.INTRODUCTION

Phishing is an illegal method which involves social and technological deceptions to obtain customer identification and financial information. In our daily life, we perform the majority of our activities on virtual platforms. The use of a computer and the internet in various fields makes our business and personal life easy. It helps us to finalize our transaction and activities rapidly in fields like trade, health, education, communication, banking, aviation, research, engineering, entertainment, and public services. The people who must be able to connect to a local network have been able to connect to the Internet anywhere, anytime with the invention of wireless and mobile technologies. This state, which offers tremendous convenience, has opened up serious loopholes from the point of view of information security. Accordingly, the requirement for users on the Internet to implement measures against potential cyber-attacks has arose. These attacks are primarily focused in the following domains: fraud, forgery, force, shakedown, hacking, service blocking, malware applications, illegal digital contents and social engineering. Based on Kaspersky's statistics, the cost of an attack in 2019 (depending on the attack size) ranges from $ 108K to $ 1.4 billion. Moreover, the amount invested in global security products and services is approximately $ 124 billion. Among these assaults, the most prevalent and most important one is "phishing attacks". It results in pecuniary loss and intangible damages.

In United States companies, they lose US$2billion annually because their customers become victim to phishing. In 3rd Microsoft Computing Safer Index Report published in February 2014, it estimated that the total global effect of phishing could reach as much as $5 billion annually. Phishing assaults are becoming effective because lack of awareness of users. As phishing attack takes advantage of vulnerabilities identified in users, it is very hard to prevent them but very essential to improve the techniques to detect phishing. The universal approach to identify phishing websites by making blacklisted URLs, Internet Protocol (IP) to the antivirus database which is otherwise referred to as "blacklist" method. To avoid blacklists attackers employ innovative methods to deceive users by altering the URL to look valid through obfuscation and various other easy technique

2. Body of Paper

2.1 Literature Review

In this work, we provide an intelligent phishing website detection system. The system is a form of an added functionality to an internet browser in the form of an automatic extension that advises the user that it has spotted a phishing site. The system draws its design upon a machine learning approach, more specifically supervised learning. We have used the Random Forest methodology because of its good performance when doing classifications. Our main concern is to pursue a better classifier performance by learning the characteristics of phishing website and select the better set of them to train the classifier. Therefore, we end our paper with good accuracy and combination of 26 features.

2.2Machine Learning based Phishing Detection from URLs

In this paper, they have suggested a phishing detection system based on machine learning by applying eight different algorithms to examine the URLs, and three different datasets in order to compare the results with other research. The experimental results illustrate that the proposed models perform exceptionally well with a success rate. In this paper, we tried to suggest a phishing detection system using some machine learning algorithms. The proposed are validated with some recent datasets in the literature and achieved results are compared with the latest works in the literature. Comparison results indicate that the proposed systems
improve the efficiency of phishing detection and achieve very high accuracy rates. In future works, first of all, it is targeted to develop a new and enormous dataset for URL based Phishing Detection Systems. It will improve our system by using some hybrid algorithms, and also deep learning model.

2.3 Detection and Prevention of Phishing Websites

The author has explained three methods of detecting phishing sites in this paper.

First is through examining different features of URL, **second** is by verifying legitimacy of website by being aware of where the website is hosted and who are taking care of it, and the **third** method is visual appearance based examination for verification of website authenticity. The writers have utilized Machine Learning methods and algorithms for assessment of these various characteristics of URL and websites

Comments: One specific challenge in this area is that criminals are continually creating new techniques to neutralize our defense strategies. In order to succeed in this regard, we require algorithms which consistently evolve in light of fresh instances and aspects of phishing URL's

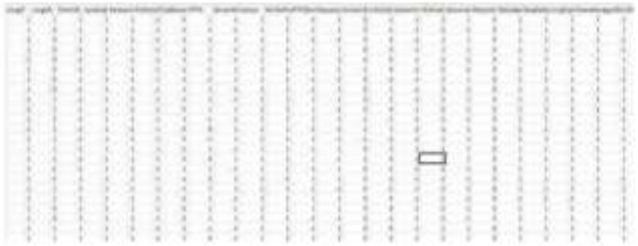

Fig 2.1 Sample of dataset

3. SYSTEM REQUIREMENTS AND SPECIFICATION

3.1 HARDWARE REQUIREMENTS

Hardware requirements refer to the essential physical components and specifications necessary for the proper functioning of a software application. These include the type and speed of the CPU, RAM and the capacity and speed of the storage devices, whether hard drives or solid- state drives. Graphics processing units (GPUs) may be specified for applications with graphical demands. These specifications guide users and system administrators in configuring and maintaining the hardware environment for seamless software operation

1. Operating System : Microsoft Windows 10/8/7
2. Hard Disk : 500 GB. 3. Ram : 4 GB.
4. Processor : Minimum Core i5
5. Laptop system with above configuration or higher level

3.2 SOFTWARE REQUIREMENTS

Software requirements encompass the essential elements necessary for the development, implementation, and functioning of a software system. These typically include the specification of programming languages, frameworks, and libraries required for development, as well as the need for specific databases or data storage solutions. Overall, software requirements serve as a comprehensive guideline, outlining the technological, functional, and operational prerequisites vital for the successful deployment and performance of a software application.

1. Operating system: Windows XP / 7
2. Coding Language: Python, HTML
3. Version: Python 3.6.8
4. IDE: Python 3.6.8 IDE
5. ML Packages: NumPy, Pandas,Seaborn, Flask, PymySql,
6. ML Libraries: whois, xgboost, favicon, beautiful soup, googlesearch.
7. ML Algorithms: Logistic Regression, Random Forest Classifier, K-Nearest neighbor, Artificial Neural Networks and XGB Classifier**.**

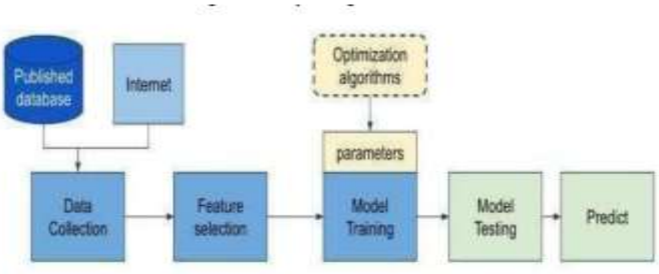

Fig -2.2 Phishing Website Detection using Machine Learning

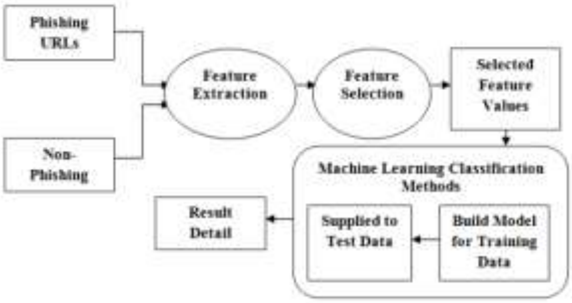

Fig 2.3 Project Flow

4. IMPLEMENTATION

In the context of phishing website detection, machine learning is a pivotal tool employed through a systematic process. It begins with the compilation of a labeled dataset, encompassing both phishing and legitimate websites. Features, such as URL structure and content analysis, are then extracted from this dataset. Following data preprocessing, an appropriate machine learning algorithm is selected, and the model is trained to recognize patterns distinguishing between malicious and genuine websites. Validation and hyperparameter tuning ensure the model's efficacy, with evaluation metrics like accuracy and precision guiding the optimization process. Once validated, the model is deployed for realtime detection, often integrated into web browsers or email clients. Continuous monitoring and updates are crucial, given the evolving nature of phishing techniques, and measures are taken to enhance the model's robustness against adversarial attacks. The integration of the machine learning model into broader cybersecurity systems provides a multi-layered defense against phishing threats. This comprehensive approach, combining machine learning with other security measures, strengthens the overall security posture and reduces the risk of falling victim to phishing attacks**.**

4.1 TYPES OF CLASSIFIERS

4.1.1 LOGISTIC REGRESSION

Logistic Regression, despite its name, is a versatile algorithm not limited to binary classification; it can be extended for multi-class text classification tasks. In this context, it works by modeling the relationship between the input features (word frequencies in the case of text) and the probability of a document belonging to each class using the softmax function. The model estimates a separate probability for each class, and the class with the highest probability is assigned as the final prediction.

4.1.2 SUPPORT VECTOR MACHINE
Support Vector Machines (SVMs) are powerful classifiers widely applied to multi-class text classification tasks. SVMs operate by finding an optimal hyperplane in a high-dimensional space that best separates the data points corresponding to different classes. In the context of text classification, each feature represents the frequency of a word in a document, and the SVM seeks to create a decision boundary that maximizes the margin between different classes.

4.1.3 K NEAREST NEIGHBOR CLASSIFIER
K-Nearest Neighbors (KNN) is one of the simplest algorithms used in machine learning for regression and classification problems which is non-parametric and lazy. In KNN there is no need for an assumption for the underlying data distribution. KNN algorithm uses feature similarity to predict the values of new datapoints which means that the new datapoint will be assigned a value based on how closely it matches the points in the training set. The similarity between records can be measured in many different ways. Once the neighbors are discovered, the summary prediction can be made by returning the most common outcome or taking the average. As such, KNN can be used for classification or regression problems. There is no model to speak of other than holding the entire training dataset

4.1.4 NAIVE BAYES CLASSIFIER
The Naive Bayes classifier is a probabilistic machine learning algorithm based on Bayes' theorem. It assumes that features are independent given the class label, hence the "naive" designation. Widely used in text classification (e.g., spam detection), it calculates the probability of each class given input features and assigns the class with the highest probability as the prediction. Despite its simplicity a nd the independence assumption, Naive Bayes often performs well in practice and is computationally efficient. It comes in different variants, such as Multinomial, Gaussian, and Bernoulli, suitable for various types of data.

4.2 CODE FOR IMPLEMENTATION OF ALGORITHMS
Code For Importing Libraries

```
from sklearn.model_selection import train_test_split#
Machine Learning Models
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import GradientBoostingClassifier
from catboost import CatBoostClassifier
from sklearn.neural_network import MLPClassifier
```

Code For Training Dataset

```
model1=LogisticRegression()
model2=RandomForestClassifier(random_state=42,max_depth=15,n_estimators=200,min_samples_split=2)
model3=XGBClassifier(n_estimators=500)
model4=KNeighborsClassifier(n_neighbors=7)
model5=DecisionTreeClassifier()
model6=CatBoostClassifier(learning_rate=0.1)
model7=SVC(kernel='linear',gamma='scale')
model8=MLPClassifier()
model9=GradientBoostingClassifier(max_depth=n,learning_rate = 0.7)
model1.fit(X_train, y_train)
model2.fit(X_train,y_train)
model3.fit(X_train,y_train)
model4.fit(X_train, y_train)
 model5.fit(X_train, y_train)
```

5. CONCLUSIONS

It is discovered that phishing attacks are extremely important and it is critical for us to obtain a mechanism to identify it. Since extremely important and personalized information of the user can leak through phishing websites, so it becomes very important to work on this task. This problem can be handled easily by applying any of the machine learning algorithm along with the classifier. We already have classifiers which provides good prediction rate of the phishing besides, but after our survey that it would be better to implement a hybrid approach for the prediction and further enhance the accuracy prediction rate of phishing websites. We have observed that existing system provides less accuracy so we suggested a new phishing method that uses URL basedfeatures and also, we trained classifiers through various machine learning. We have obtained the results for which the site is to be tested whether it is phishing or not using five classifiers. The project also has other phishing variants such as smishing, vishing, etc. to make the system complete. Peering even farther ahead, the methodology must be assessed on how it would address collection growth. The collections will hopefully grow incrementally over time so that there will be a means to incrementally apply a classifier to the new data.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our Head of the Department, **Dr. B Sreepathi** and guide, Prof. **Soumya A**, for their continuous support, valuable guidance, and encouragement throughout the development of this research project. Their insights and constructive feedback played a vital role in enhancing the quality of this work.We are also thankful to the Department of Information Science and Engineering,Rao Bahadur Y Mahabaleshwarappa Engineering College,for providing the necessary resources and a collaborative environment that helped us explore and apply advanced concepts in Natural Language Processing and machine learning. A special thanks to our peers and friends who participated in the testing phase of the YT Sentiment Explorer application and provided useful feedback to improve its usability.Finally, we thank our families for their unwavering support and motivation throughout this journey.

REFERENCES

1. V. Patil, P. Thakkar, C. Shah, T. Bhat and S. P. Godse, "Detection and Prevention of Phishing Websites Using Machine Learning Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

2. M. H. Alkawaz, S. J. Steven and A. I. Hajamydeen, "Detecting Phishing Website Using Machine Learning," 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), 2020.

3. J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 2020.

4. M. M. Vilas, K. P. Ghansham, S. P. Jaypralash and P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019.

BIOGRAPHIES

Mrs. Soumya A

Assistant Professor,
Dept of ISE,
RYMEC, Ballari.

Chandrashekar KM

Final Year student of Information Science and Engineering (ISE) Department, RYMEC, Ballari.

Prajwal

Final Year student of Information Science and Engineering (ISE) Department, RYMEC, Ballari.

Vijay Kumar S Biradar

Final Year student of Information Science and Engineering (ISE) Department, RYMEC, Ballari.

D Prashanth Reddy

Final Year student of Information Science and Engineering (ISE) Department, RYMEC, Ballari.