Phishing Website Detection Using Machine Learning

Vaishnavi Indurkar, Shruti Jagtap, Tejasvi Khote, Dhanashri Mane Padmabhooshan Vasantdada Patil Institute of Technology, Pune,India Department of Computer Engineering

Prof. Soniya Waghmare

Abstract

Phishing Website URL data has become a severe issue in today's society, along with the rapid rise of internet users. People are sending URLs with dangerous links that may damage our system and your system and engage in unlawful and unethical behavior, phishing, fraud, and other activities. The process of setting up a fake profile and URLs is much simpler for Phishing. They prey on gullible victims of these scams by posing as real people in their phishing website URL. Therefore, it's important to recognize Phishing Website URL data. In this research, machine learning methods will be used to detect Phishing Website URL data that are false. The most effective algorithm for Phishing Website URL detection will be determined after discussing machine learning methods and applying them to our data sets. The proposed research detects Phishing Website URLs using a mixed machine-learning approach. The system initially uses natural language processing (NLP) to extract features and train the module. With machine-learning hybrid models, a model for Phishing Website URLs detection is created in the current study. The Phishing Website URL dataset was subjected to various machine learning techniques: Multilayer Perceptron's (ANN), NB, RF and SVM. A comparison study is carried out to verify the effectiveness and performance of the suggested approach.

Keywords— phishing detection, Phishing URL's, Cyber Attack, Machine Learning classifiers.

1. INTRODUCTION

Phishing is the most commonly used social engineering and cyber-attack. Through such attacks, the phisher targets naïve online users by tricking
`them into revealing confidential information, with the purpose of using it fraudulently. In order to avoid getting phished, users have awareness of phishing websites. Have a blacklist black list of phishing websites which requires the knowledge of website being detected as phishing. Detect them in their early appearance using machine learning algorithms. Of the above three, machine learning based method is proven to be most effective than the other methods. Even Nevertheless, those who utilize the internet continue to divulge private information to phishing websites. The system that detects phishing website URLs can be thought of as a classification problem that is divided into two groups: phished and normal. In the artificial intelligence discipline of machine learning, a system is given the capacity to learn without explicit programming. We implement supervised machine learning algorithms for categorization in our model. Using existing examples as a basis, supervised learning algorithms make predictions about the nature of unknown data. These algorithms are a subset of algorithms used in machine learning, which learn from data iteratively.

URL is slightly changed, but it's not the . Filling in this form would give the attacker full access to the victim's Gmail account. The kind of theft and fraud that could take place by just acquiring the details of someone's or some organizations' account couldn't really be imagined. All other account are controlled by the Gmail account. That could be a huge threat. Microsoft Outlook fraud is the second-most targeted and

Google drive being the third. Other targets are Facebook, bank logins and Paytm, PayPal etc.

2. Literature Survey

According to [1] the approach incorporates machine learning, optical character recognition, and natural language processing. Optical character recognition extracts text from pictures, while natural language processing distinguishes between standard text and slang by detecting and classifying the language. The bag-of-words model is then used to remove the characteristics from the chosen photos, after which the machine learning method is used to look for any potential spam. The program can also forecast if the picture includes any dangerous elements.

According to [2] several studies have examined how spam activity affects finance, the economy, marketing, business, and management. Other research has concentrated on examining how spam affects security and privacy. As a result, several anti- spam techniques are provided in the literature that prevents or filter spam emails the challenges of the current anti-spam practices, identifying current issues, and implementing a better anti-spam model. A novel Multi-Natural Language Anti-Spam (MNLAS) paradigm based on agents is suggested in this respect. Both visual information, such as photos & sentences in both English and Arabic, is processed by the MNLAS model during the spam filtering process of an email. A method for detecting spam emails that makes use of machine learning algorithms enhanced by bio-inspired methodologies has been reported in

According to [3] A review of the literature is done to look into efficient methods that are used to different datasets to get great results. Using seven different email datasets, a thorough investigation was carried out using feature extraction and pre-processing to create machine learning models using Naive Bayes, Support Vector Machine, Random Forest, Decision Tree, and Multi-Layer Perceptron. Two examples of bio-inspired algorithms that are used to improve classifier performance are genetic algorithms and particle swarm optimization. The Multinomial Nave Bayes with Genetic Algorithm produced the best overall results. Using a few common algorithms.

According to [5] the same system can distinguish between spam and legitimate emails. For efficient outcomes, our suggested technique creates a dictionary of features & trains them using machine learning. Unsolicited bulk email (UBE) and unsolicited commercial email (spam) are terms used on the internet (UBE). Internet users often refer to emails of this kind forwarded for commercial or company advertising objectives as trash emails. Currently, spam emails are growing daily and causing problems for users. By using a spam detector, we can determine whether a message is spam, enhancing user efficiency.
According to [6] they get reliable findings there used multiple classification & regression methods to datasets based on spam and unsolicited bulk email, where they found spear phishing techniques. To get better results despite current algorithms, we presented ensemble approaches in this research, such as boosting, bagging, stacking, and voting, for increased classification accuracy and efficiency.

According to [7] using WEKA with our email management system, which uses the PHP-ml package that GitHub offers, we assess machine learning algorithms on spam email detection. The results reveal that the naive Bayes algorithm delivers reasonable accuracy and precision. Additionally, a comparison between SVM and the earlier method is made regarding the accuracy and the dataset employed. They thoroughly understand the most acceptable text categorization method, the naive Bayes algorithm. Its mathematics and pseudocode are described, and the choice of several characteristics for more sophisticated outcomes is covered.

According to [8] analyses of several machine learning methods, including Adaboost & Stochastic Gradient Descent, to filter spam emails (SGD). For the pre-processing step, they used R. The classifiers were built in

the Orange program using Adaboost and SGD. Adaboost and stochastic gradient descent (SGD) algorithms generated actual positive values of 100% and 98.1%, respectively, and false favorable rates of 0.0% and 1.9%, according to experimental findings using the Orange tool. These algorithms are among the most acceptable options for spam filtering techniques due to their high accuracy and beneficial outcomes.

According to [9] a spam detection model based on support vector machines is put forward while carefully looking for the best settings to improve performance. Experimental findings demonstrate that the suggested model outperformed all previously presented models on the same well- known dataset used in this study. With the help of a well-liked and often-used standard database, a better email spam detector based on an SVM classifier has been suggested, trained, and evaluated. According to empirical simulation findings, the told SVM-based technique beat other previously published spam- detecting schemes assessed on the same widely used database utilized in this work.

According to [10] has discovered some performance measures are often used to gauge how well the review spam detection algorithms operate. Finally, a general overview of various feature extraction techniques from review datasets, a suggested taxonomy for spam review detection techniques, assessment metrics, and publicly accessible review datasets are presented in this study. Websites are becoming the primary medium for people to express themselves. People may quickly express their opinions on goods and services through e-commerce websites, forums, and blogs.

Before making a purchase, most consumers check reviews of products and services. Today, everyone on the internet recognizes the value of these online evaluations for merchants and other consumers. Vendors might use these assessments to inform the creation of their further marketing plans.

3. System Feature

Data Collection

We gather data from a variety of sources, including UCI ML repository, Kaggle and several real-time data sources. Prior to executing the classification activity, the data must be pre- processed in order to enhance the outcome by addressing the missing values and removing the redundant features contained in the chosen dataset. In order to get the greatest results during the Data Mining process, the dataset must be treated swiftly.

Pre-processing and Normalization: There may be a lot of useless information and gaps in the data. Data preparation is carried out to handle this portion. Numerous data pre-processing techniques, such as data cleansing, data transformation, and data reduction, have been utilized at this step.

Data Cleaning: Data cleaning deals with issues including missing information and noisy data. When there are gaps in the information, many approaches have been taken, such as ignoring the tuples or filling in the gaps. Machines cannot understand null values, which can be present in data. This noisy data could be the consequence of inadequate data collection, inaccurate data entry, etc. To address it, the binning method, grouping, and regression are employed.

Data transformation: Is the process of transforming data into a format that is appropriate for the mining process. This method involves discretization, attribute selection, and normalization. Then, we will use a variety of preprocessing techniques, including lexical analysis, stop word removal, stemming (Porters method), index term selection, and data cleaning

Selection and extraction of features:

This process retrieves a range of features from the input data. A feature selection threshold is then used to

the extracted features to normalize them and remove any redundant or superfluous characteristics for training. Many hybrid properties are extracted from the normalized data with relational features, and training is done by choosing an optimization approach. Feature selection from fully extracted features has been done using the hybrid method; choosing the highest quality boosts the classification accuracy. During the feature extraction process, a lot of unnecessary characteristics emerge. These must be removed when selecting the relevant portions. This method's advantage is that it offers a suitable feature selection for each unique feature set.

Classification: Following the module's successful execution, the training module receives the chosen characteristics as input, producing thorough Background Knowledge for the system as a whole. Once the training model is ready, we can feed it the testing data to obtain the classification prediction. Preprocessing, vectorization, and classification of the testing text are all part of the testing stage. Using machine learning techniques, the module testing assesses the predicted performance of the system. In this step, several datasets were used to evaluate the system's performance.

4. Use of the SDLC Model

Every project activity is carried out and tracked using this plan as a guide. It is going to be used for the duration of the project and updated to reflect real project accomplishments and plans. The software project plan is displayed in Fig. 1.1. The following was the plan and requirements collection for the project's first phase:

We must comprehend the problem definition and system reliability during requirement collecting and analysis of the suggested system. It also entails learning about the necessary hardware and software. We must create a rough design of the project's overall work flow during the designing phase, which will provide further details.

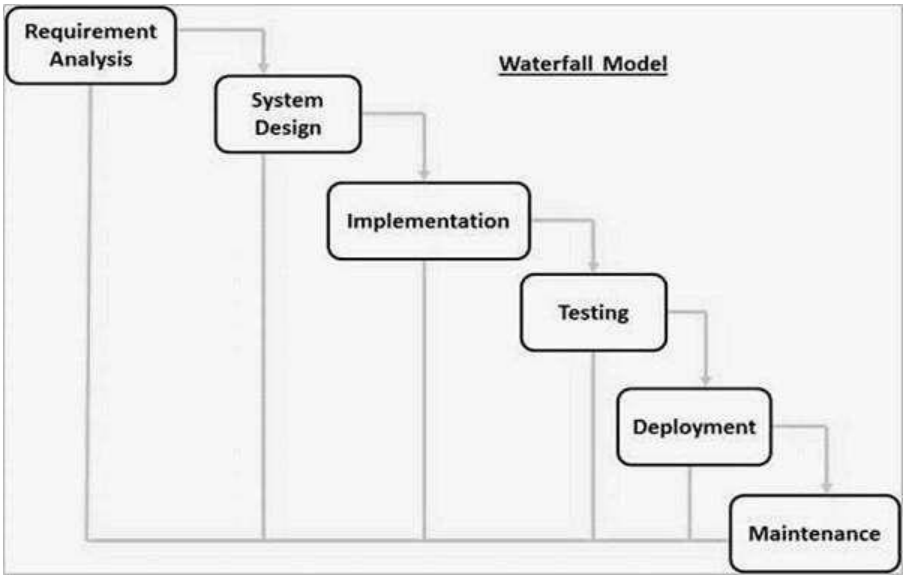


Fig Project plan

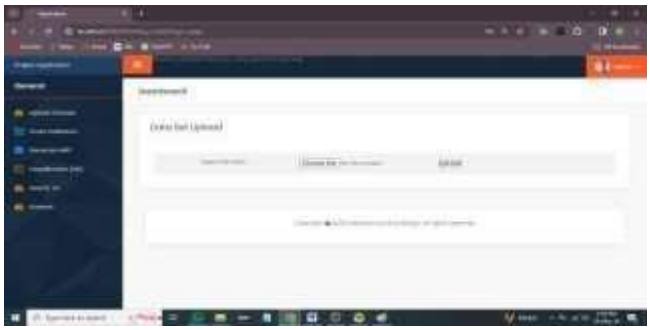
Last but not least part of development is testing. There are many testing approaches are available. Testing will help to decide whether our system fulfills our requirements or not. In chapter 6 we explain some parameters of

testing. If testing gives proper result it will go for user acceptance. If any change occurs, the whole plan will get repeated.

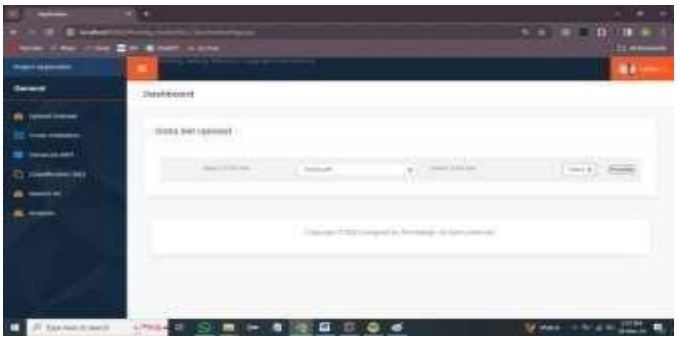
In this system, waterfall model is being adapted for software development and testing. System acceptance testing followed by user acceptance testing is being used as sign off for validity of the developed system.

1. **Login Page**

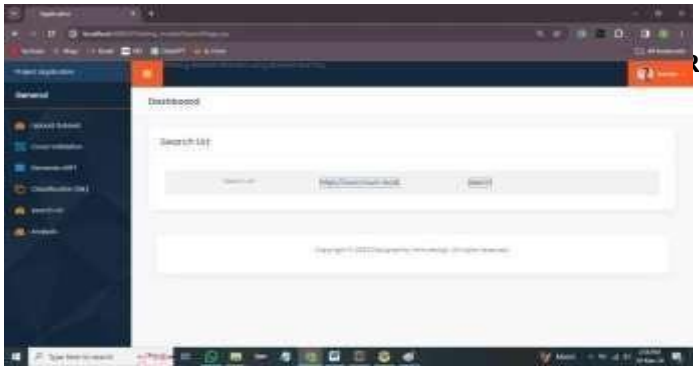
2. **File Dataset Upload**



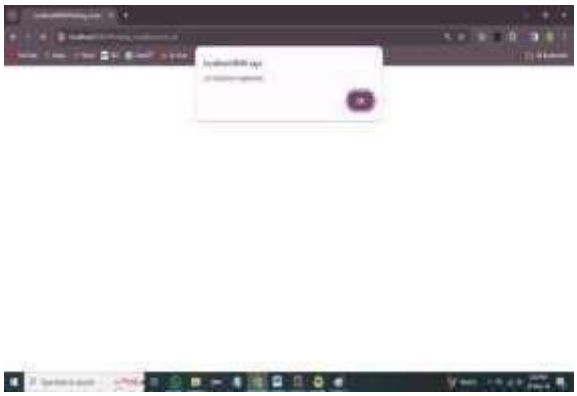
3. **Prediction**



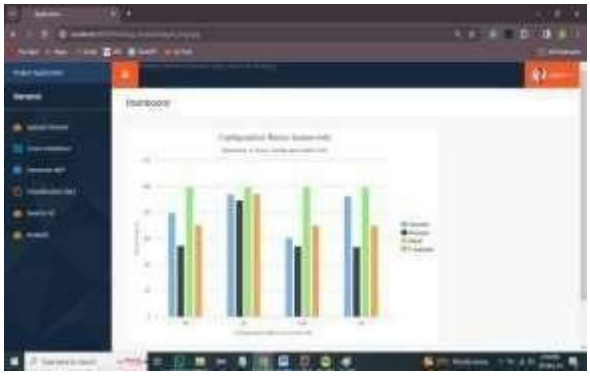
4. **Search**



5. Result



6. Analysis



5. Architectural Design

Apply preprocessing in the training and testing phase, and then proceed with features extraction and selection. Train the system using hybrid approaches (machine learning) algorithm, and generate training rules. Classify all test data and normal as well as Phishing based on calculator weight for each test sample. Finally, predict the accuracy of the entire system using various confusion matrixes and provide the analysis accuracy with the true positive and false negative of the system.

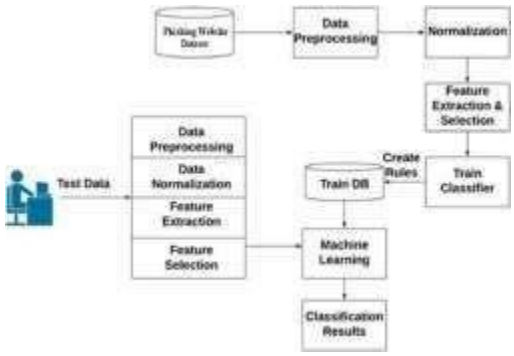


Figure : System Architecture

6. Mathematical Model

Let S is the Whole System Consist of S= {I, P, D, O}
I = Input Phishing Website data. P = Process:
D = Dataset
Step1: User will enter the query.
Step2: After entering querythe following
Operations will be performed. Step3: Data Preprocessing.
Step4: Feature extraction and feature selection. Step5: Training and Testing dataset.
Step6: Classification.
Step7: Final output optimized classifier and its
performance indicator.
O= Output (Predicted Values)

8. Conclusions

The samples of data sources and data collecting structures have led to a large increase in the data available for cyber security experts. To process such large volumes of data, scalable massive data processing solutions are needed. This review investigates how the system works to detect malicious content from incoming Phishing Website URLs using natural language processing and machine learning algorithms. The detect such entries, the system needs to analyze the entire metadata of the system, and according to the selected features, it built a training module. Different techniques are introduced in the proposed review for supervised learning, and detection analysis has been done for the machine learning algorithm. Implementing machine learning algorithms on synthetic and real-time phishing website is an interesting future direction for this research.

References:

1Yaseen, Yaseen Khather, Alaa Khudhair Abbas, and Ahmed M. Sana. "Image spam detection using machine learning and natural language processing." Journal of Southwest Jiaotong University 55.2 (2020).
2 Mohammed, Mazin Abed, et al. "An anti-spam 3detection model for emails of multi-natural language." Journal of Southwest Jiaotong University
54.3 (2019).
4 Gibson, Simran, et al. "Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms." IEEE Access 8 (2020):
187914-187932.
5Nandhini, S., and Jeen Marseline KS. "Performance evaluation of machine learning algorithms for email spam detection." 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic- ETITE). IEEE, 2020.
6 Govil, Nikhil, et al. "A machine learning-based spam detection mechanism." 2020 Fourth International Conference on Computing
Methodologies and Communication (ICCMC). IEEE, 2020.
7 Chandra, J. Vijaya, Narasimham Challa, and Sai
Kiran Pasupuletti. "Machine learning framework to analyze against spear phishing." Int. J. Innov. Technol. Exploring Eng. (IJITEE) 8 (2019): 12.
8 Bibi, Asma, et al. "Spam mail scanning using a machine learning algorithm." J. Comput. 15.2
(2020): 73-84.