Efficient Email phishing detection using Machine learning

ABSTRACT

Emails are frequently utilized as a way of personal and professional communication. Banking information, credit reports, login data, and other sensitive personal information are commonly transmitted over email. This makes them valuable to cybercriminals, who can exploit the knowledge for their gain. Phishing is a technique used by con artists to steal sensitive information from people by impersonating well-known sources. The sender of a phished email can persuade you to disclose personal information under pretenses. The detection of a phished email is treated as a classification problem in this research, and this paper shows how machine learning methods are used to categorize emails as phished or not. SVM classifier attains a maximum accuracy of 0.998 percent in email classification.

**EXISTING SYSTEM**

For identifying legal and fraudulent web pages, is-based phishing detection systems use two lists: whitelists and blacklists. Phishing detection systems that use whitelists create secure and genuine websites that deliver relevant information. Every website that isn't on the whitelist is regarded as potentially dangerous.[5] built a system that creates a whitelist by logging the IP address of each site that the user has visited with a Login user interface. When a user accesses a website, the system will alert them if its registered information is incompatible.

The authors of [15] classified phishing websites using URL parameters such as length, number of unique characters, directory, domain name, and file name to identify them. The system uses support Vector Machines to classify websites that are not online. Adaptive Regularization of Weights, Confidence Weighted, and Online Perceptron are utilized for online classification. According to the trials ' findings, using the Adaptive Regularization of Weights algorithm improves accuracy while reducing system resource requirements.

Authors in [16]used a nonlinear regression technique to detect whether a website is phishing or not in a recent study. They train the system using harmony search and support vector machine meta-heuristic techniques. Harmony search, they claim, has a higher accuracy rate of 94.13 percent and 92.80 percent for train and test procedures, respectively, thanks to the use of around 11,000 web pages.

In [17] created a phishing detection system that uses adaptive self-structuring neural networks to classify the data. It has 17 features, some of which are reliant on third-party services. As a result, real-time execution takes substantially longer; yet, it can achieve higher accuracy rates. It only has 1400 items in its dataset, yet it has a reasonable acceptance rate for noisy data.

Yank in [18] provides an anti-phishing strategy that employs machine learning to identify phishing websites from legal ones by extracting 19 features from the client side. They used PhishTank (2018) and Openfish (2018) phishing pages and 1918 authentic web pages from Alexa popular websites, online payment gateways, and prominent banking websites. Their proposed approach achieved a 99.39 percent true positive rate using machine learning[4].

**Disadvantages**

* An existing system not implemented an effective ML Classifiers like SVM,RF,NB.
* An existing system not implemented for large number of datasets.

Proposed System

The attackers add subdomains to the links to make them appear authentic. The number of dots in the link rose as subdomains were added. As suggested by In a valid email, the number dots should not be used. More than three [three]. This is a binary feature, meaning it determines whether or not a link exists. It would be in the mail if the number of dots was more prominent than three. This is a phished email. The total number of links is: In general, phishing emails provide more information. In comparison to ham, the transmitter attempts to send many links. By tricking the user, you might direct him to an illicit website. This is a recurring feature.

The presence of JavaScript in an email indicates that the sender is either trying to conceal information or activate specific browser changes [18]. This is a one-of-a-kind feature. The presence of the script> tag in an email indicates that it has been phished. Form tag: Phishing emails feature forms integrated into them to acquire information from users. This is a binary characteristic, meaning that the presence of a form tag indicates that the email is phished.

HTML emails allow the sender to include embedded graphics and URLs, which are not possible with plain text emails. If the email has an HTML tag, it is considered

phishing. This is a one-of-a-kind feature. The use of action words in emails shows if the sender expects the recipient to do a specific action, such as clicking on a link, filling out a form, or submitting detailed information. This is a recurring feature.

The word PayPal indicates that the sender is posing as a member of a recognized organization. The word "PayPal" appears in the mail's links or the "from" section, implying that the sender is affiliated with PayPal. This is a one-of-a-kind feature.

The presence of the term bank is a binary indicator indicating the message is about banking. Either the sender is posing as a member of the financial organization, or the reader is looking at the reader's credentials. The word account appears in the email, indicating that it seeks emails tied to an account. It could be a social media account, a bank account, or something else entirely. It's a one of- a-kind feature.

**Advantages**

* SVM is a supervised technique often used for text categorization because of its speed and accuracy. It generates a hyper\_plane, a two-dimensional line that best separates the categories, based on the training data. The decision boundary is the name given to this hyper\_plane
* The naive Bayes classifier[20] is a probabilistic technique that uses the Bayes theorem to classify sample data.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).