IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006303 |

# Phising Email Detection

**Wakchaue Rohan[1], Kate Rushikesh[2], Torpe Aditya[3], Sayyad Imran Majid[4]**

**Prof. Salunke M.D.[5]**

Student, Dept. of Comp, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Ahmednagar, India [1]

Student, Dept. of Comp, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Ahmednagar, India [2]

Student, Dept. of Comp, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Ahmednagar, India [3]

Student, Dept. of Comp, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Ahmednagar, India [4]

Guide, Dept. of Comp, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Ahmednagar, India [5]

**ABSTRACT**: Phishing is a type of extensive fraud that happens when a malicious mails act like a real one keeping in mind that the end goal to obtain touchy data. In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavors in messages and identifying phishing substance on sites. However, detecting phishing Mails is a challenging task, as most of these techniques are not able to make an accurate decision dynamically as to whether the Arrived mail is spam or not. Propose system will compare mail with spam keyword database ,if content in mail matched with database contents then mail is detected as spam. saving.

**KEYWORDS**: NLP, Prediction, Spam words, Phishing.

## I. INTRODUCTION

Phishing is a term used to describe a malicious individual or group of individual who scam users. for example account points of interest, passwords or MasterCard numbers. The fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavors in messages and identifying phishing substance on mails, phishers think that new and half breed strategies to go around the accessible programming and the systems. Phishing is trickery system that uses a blend of social designing more, innovation to assemble delicate and individual data, for example, passwords and charge card sub tile elements by taking on the appearance of dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to originating from honest to goodness sources like money related foundations, ecommerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email. Phishing is a fraudulent attempt to tricking people into giving important information like usernames and passwords, credit card details, sensitive bank information, etc., by way of instant messaging, email spoofing or using fake mails whose look and feel gives the appearance of legitimate mails. System will decides whether a mail is phishing or normal thanks to the Bayesian classification algorithm and the scores added to the database. It is a instantly perceived as a spam mails by the words that are exciting, phrases that increase the desire for shopping, and which contain the unwanted content in the mail.In this paper, we extract 18 features including header-based features, URL-based features, scriptbased features and psychological features. In order to extract the features, we analyse the email header structure, the URL information and the function of email-script. Furthermore, according to the fact thatphishing emails use recipients' psychological weakness, we proposed the email psychological features. Then, we build SVM module based on these 18 features together to train and classify the emails. This method is evaluated on a testing dataset and demonstrates a better performance in detection phishing emails. The rest of the paper is organized as follows: the section 2 describes the related work on phishing email detection; The section 3 presents the SVM classifier used in our paper and details our proposed 18 hybrid features; The section 4 describes and analyses the experimental results; The section 5 presents the conclusion and describes the future work.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006303 |

## II. RELATED WORK

Generally, phishing emails detection can be divided into two types: the blacklist method and the machine-learning method. The blacklist characteristic library consists the sender blacklist and the URL blacklist. If some blacklist words in emails title and emails content, the email can be regarded as a phishing email. Although blacklist is simple and efficiency, they failed to detect new phishing attacks. Meanwhile, blacklist collection is time-consuming. The machine-learning method is designed to classify new phishing emails. These methods have the highest detection precision and efficiency among the existing phishing email detection methods. In 2006, Ian Fette [4] et al. proposed a machine learning-based phishing email detection method called PILFER. This method extracted 10 features from emails including URL-based and script-based features, and finally gained high accuracy 96%. However, with the rapidly change of phishing email, some features in this method are not the best indicators now. Sunil B. Rathod [5] et al. applying Bayesian method for detecting legitimate and phishing emails employing supervised learning across features extracted. The experimental results demonstrated that using the Bayesian classifier can achieve an accuracy over 96.46% in detecting the real world email datasets. Carthy [6] et al. extracted 40 features from URL-based features, script-based features, subject-based features, body-based features and sender-based features to classify phishing emails. This paper calculated the information gain and entropy to evaluate the effectiveness of their features. The results of experiments showed that URL and script features were the most effective features among 40 extracted features to detect phishing emails. Xiang [7] et al. proposed the detection method of "CANTINA+" based on the network analysis tool "CANTINA". This method extracted 8 new features by using HTML, DOM, search engine and thirdparty services, and then used heuristic rules to filter out the website without login box. They employed the extracted 15 features such as URL lexical features, Form features, WHOIS information, PageRank value, etc. to train and test their classifier model. The experimental results also gave a satisfactory performance than the previously detecting method. Naghmeh [8] et al. employed word embedding or vectorisation and proposed a Neural Network (NN)-based model. They used publicly available email datasets that contains legitimate and phishing emails for detection and classification of phishing emails. The experimental results showed their detection module of the phishing emails offered a satisfactory performance in terms of accuracy, TPR, FPR, network performance and error histogram.

## III. METHODOLOGY

In this section, we detail the SVM classifier used in our paper and our proposed 18 hybrid features we extracted. The architecture of proposed phishing emails detection is shown in Figure 1. We extracted 18 hybrid features from four parts including email-header structure, email-URL information, email-script function and email psychological features. Then we choose SVM classifier to detect phishing emails.

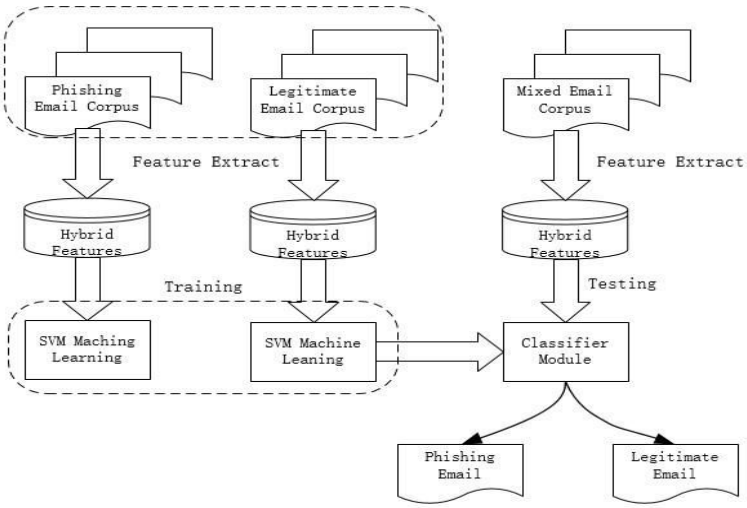


*Figure 1.*Architecture of Proposed Method

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006303 |

*1. Support Vector Machine*

Support Vector Machine (SVM) is a supervised learning model for analyzing data in classification and regression analysis. The previously research has shown SVM is effective to detect phishing email.

We denote the i-th mail feature vector by , and using to represent the label of i-th mail. If the mail is legitimate mail, equals 1. If the mail is phishing mail, is -1. Then the total feature set of n emails is:

$$D=(x_1,y_1),(x_2,y_2) \ldots (Xn,Yn) \quad (1)$$

We use available phishing and legitimate mail datasets to optimize the parameters of SVM, and use the classifier composed of the optimized parameters for the detection of test set samples. The support vector machine model used in this paper is shown in figure 2.

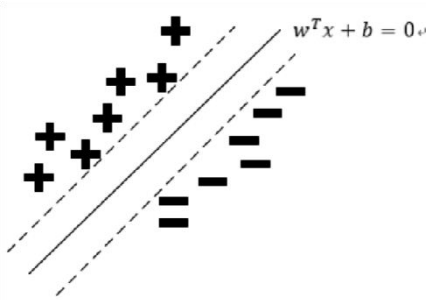


*Figure 2. Support Vector Machine*

The model generates a hyperplane which can maximize the classification of data point. The hyperplane can be represented as:

$W^T x + b \quad (2)$

Where all this w and constant b stuff describes the hyperplane for our data. To find the distance from to the hyperplane, we must measure perpendicular to the line. This is given by:

$$|w^T x + b| / ||W|| \quad (3)$$

To judge the correctness of classification results. Then, this paper uses the training feature set D to optimize the parameters and the constant b of SVM, so that the SVM composed of the optimized parameters can correctly classify the training data. Finally, this work calculates the classification error and accuracy about testing data using the current training classifier.

*2. Feature Extraction*

This section details our proposed 18 hybrid features for phishing emails detection. The hybrid features consist of four email header-based features, nine email URL-based features, four email script-based features and an email psychological feature. Let be the i-th feature of a mail, so the n-th mail's feature vector can be represented as = ( $_{1\,2\ldots18}$). The extracted features in this paper is shown in table I.

*Table 1. Feature List*

| Feature | Description |
|---|---|
| 1 | Blacklist words in title, such as bank, pay, account, password |
| 2 | Inconsistency between replies address and send's address |
| 3 | Inconsistency between sender domain and Message-Id domain |
| 4 | Whether email is html format |
| 5 | Whether presence of IP address in URL |
| 6 | Whether presence of image links |
| 7 | Whether presence of abnormal ports in URL |
| 8 | The number of "%" character in URL |
| 9 | The number of "@" symbol in URL |
| 10 | The number of "." in URL |

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006303 |

| | |
|---|---|
| 11 | The number of URLs in email |
| 12 | The number of domains |
| 13 | Inconsistency between where the URL actually leads to and the text the link appeared |
| 14 | Whether presence of JavaScript in email |
| 15 | Whether the JavaScript can change the status bar |
| 16 | Whether the JavaScript can change the pop-up event |
| 17 | Whether the JavaScript can change the on-Click event |
| 18 | The psychological features in email |

*2.1. Header-based Features:* The header of the email contains the basic email attribute. But the attribute is often ignored by the recipient. The phishers can modify the header attribute to achieve the purpose of phishing. This section details four header features to identify phishing emails.

(1) Blacklist words in title

In order to steal recipient information, most phishing emails are always disguised as a bank or online payment website in the title. The blacklist is collected by analyzing the titles of legitimate emails and phishing emails. Based on Lew May's work [9], we totally collected 18 blacklist words used in our work to detect the email title. The blacklist is shown in table II. If there is a blacklist keyword in the email title, let the feature value be 1, otherwise, it is set to 0.

*Table 2. Blacklist Keywords*

| | | |
|---|---|---|
| Account | Debit | Recently |
| Access | Information | Risk |
| Bank | Log | Security |
| Client | Notification | Service |
| Confirm | Password | User |
| Credit | Pay | Urgent |

(2) Inconsistent email address

An attacker can spoof the sender's address to a real official address and change the replies address to a phishing email address. The recipients usually don't pay attention to the replies' address, and carelessly return financial data to the phishing email address. If the sender's address and the replies address are inconsistent, let the feature be 1, otherwise, it is set to 0.

(3) Inconsistent domain name

The sender's domain name can be changed by the phishers, but the Message-Id domain name cannot be changed. For example, the Message-id of the email sent by the Apple is 128389078.84832701. 1542814793094. JavaMail.email@email.apple.com. The domain name of the email is email.apple.com. Generally, enterprises with their own email servers have the same sender domain name as Message-id domain name. If Message-Id domain name is not same as the sender's domain name, let the feature value be 1, otherwise, it will be 0.

(4) Whether a html email

The html function can help attackers to hide phishing links and fake information, so that the phishing email can't be easily identified by recipients. A phishing email does need to use html format. If the email is html format, the feature value will be 1, otherwise, it will be 0.

## IV. EXPERIMENTAL RESULTS

We extracted 18 hybrid features from each email (abbreviated as HF). To demonstrate the performance of our method, we compared our method with Ian Fette's approach PILFER [4] and May's approach (abbreviated as MA) [9]. The results are shown in Table 4. Our proposed method achieved overall accuracy of 95.00%, true-positive rate of 99% and false-positive rate of 9%. Compared to PILFER and HA, our method can obviously improve the

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006303 |

TPR and Accuracy of detecting phishing emails. Although the FPR of our method is higher than PILFER, it's reasonable and acceptable. Because if the detection of the phishing email is not accurate enough and strict, more phishing emails will be clicked by the recipients, which will cause recipients great financial loss and information leakage. Overall, our method has better performance in detecting phishing emails.

*Table 4.*Summary of Classification Results

| Methodology | HF | PILFER | MA |
|---|---|---|---|
| TPR | 99% | 86% | 94.5% |
| FPR | 9% | 5% | 9.5% |
| Precision | 91.7% | 94.5% | 90.8% |
| Accuracy | 95.00% | 90.5% | 92.5% |

To further demonstrate the effectiveness of the features we selected, we calculated the average value of each of our 12 binary features in both phishing and legitimate messages. The result is shown in Table 5. The features with large difference between phishing emails and legitimate emails in the average value are the number of ".", the number of domains, the average value of blacklist, the average value of whether html emails, the average value of whether contains image links and the average value of inconsistent URL. For the 5 numerical features, their average values and standard deviations per-class are shown in Table 6. These features have obvious difference in the average value between phishing emails and legitimate emails. Furthermore, these values of features in phishing emails are more stable. Our proposed features are highly effective for us to detect phishing emails.

*Table 5. Average Value of the Binary Features*

| Feature | Legitimate | Phishing |
|---|---|---|
| Blacklist | 0.0234 | 0.4989 |
| Inconsistent address | 0.1295 | 0.1930 |
| Inconsistent domains | 0.6290 | 0.8990 |
| Html emails | 0.0193 | 0.4479 |
| Ip link | 0.0 | 0.0637 |
| Image link | 0.0214 | 0.4250 |
| Abnormal port | 0.0014 | 0.0136 |
| Inconsistent URL | 0.0203 | 0.3503 |
| Contain JavaScript | 0.0105 | 0.0532 |
| Change statue-bar | 0.0 | 0.0096 |
| Change pop-up | 0.0 | 0.0396 |
| Change on-Click | 0.0 | 0.0249 |

*Table 6. Mean, standard deviation of the Numerical Features*

| Feature | | | | |
|---|---|---|---|---|
| Number of "%" | 0.223 | 1.27 | 0.115 | 0.756 |
| Number of "@" | 0.015 | 0.182 | 0.303 | 0.503 |
| Number of "." | 14.179 | 26.186 | 9.397 | 41.762 |
| Number of URLs | 4.342 | 7.570 | 3.200 | 11.733 |
| Number of domains | 2.193 | 1.703 | 1.747 | 3.482 |

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006303 |

*Table 7. Classification results of hf and 17 features*

| Methodology | HF | 17 Features |
|---|---|---|
| TPR | 99% | 97.5% |
| FPR | 9% | 9% |
| Precision | 91.7% | 91.5% |
| Accuracy | 95.00% | 94.25% |

In order to evaluate the effectiveness of psychological features, we tested the 17 features without the psychological features. The experimental results are shown in Table VII, the 17 features achieved an Accuracy of 94.25%, Precision of 91.5%, TPR of 97.5% and FPR of 9%. Compared with 18 features, its Accuracy, Precision, TPR and FPR were all decreased. The reason is that some phishing emails expressed strong emotions in order to take advantage of the psychological weakness of the recipient, resulting in more obvious psychological feature of such emails than ordinary emails. Therefore, this feature can reduce the proportion of phishing emails being recognized as normal emails.

## V. CONCLUSION AND FEATURE WORK

In this paper, we have extracted 4 header-based features, 9 URL-based features, 4 script-based features, and 1 psychological feature. The SVM has been used to train and test data set. The experimental results show that compared with the PILFER method and May's method, our method can have better performance in terms of TPR and accuracy. Although the FPR of our method is 9%, it's acceptable and reasonable. We also calculated the average of the binary features in phishing emails and legitimate emails. The results show the binary features we extracted have an obvious difference between phishing emails and legitimate emails. By using the psychological features, we can effectively reduce the proportion of phishing emails detected as legitimate emails. Overall, our proposed method has a good performance in detecting phishing emails.

In the future, we will further study the means by which an attacker users the recipient's weakness. We expect to find more effectively psychological features which can be used directly to detect the phishing emails.

## REFERENCES

[1] "Phishing activity trends report-second quarter 2016 [EB/OL]," bhttp://docs.apwg.org/ reports/ apwgtrendsreportq22016.pdf

[2] "APWG Phishing trends reports-second quarter 2018," http://www.antiphishing. org/.

[3] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg and E. Almomani, "A survey of phishing email filtering techniques", Communications Surveys Tutorials, IEEE, vol15, iss4, pp. 2070– 2090, 2013.

[4] I. Fette, N. Sadeh and A. Tomasic, "Learning to detect phishing emails", Proceedings of the 16th International Conference on World Wide Web, pp. 649–656, 2007.

[5] Sunil B. Rathod, Tareek M. Pattewar. Content Based Spam Detection in Email using Bayesian Classifier [C] // International Conference on Communications and Signal Processing (ICCSP). 2015.

[6] Toolan, Fergus and Carthy, Joe, "Feature selection for spam and phishing detection," eCrime Researchers Summit (eCrime), 2010, pp. 1–12.

[7] XIANG G, HONG J, ROSE C P, et al. Cantina+: A feature-rich machine learning framework for detecting phishing Web sites [J]. ACM Transactions on Information and System Security (TISSEC), 2011, 14 (2): 21.

[8] NaghmehMoradpoor, Benjamin Clavie and Bill Buchanan. "Employing machine learning techniques for detection and classification of phishing emails," Computing Conference, 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details