# Phishing Email Detection Based on Hybrid Features



# Phishing Email Detection Based on Hybrid Features

**Zhuorao Yang*, Chen Qiao[a], Wanling Kan[b] and Junji Qiu[c]**

School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding author e-mail: alex@bupt.edu.cn, [a]qc666@bupt.edu.cn, [b]kwl@bupt.edu.cn, [c]junji23@bupt.edu.cn

**Abstract**. As an attack of social engineering, phishing email has caused tremendous financial loss to recipients. Therefore, there is an urgent need for phishing email detection with high accuracy. In this paper, we proposed phishing emails detection based on hybrid features. By analysing the email-header structure, email-URL information, email-script function and email psychological features, we extracted 18 hybrid features. Then we chose Support Vector Machine (SVM) classifier to evaluate our experiments. Experiments are performed on a dataset consisting of 500 legitimate emails and 500 phishing emails. The proposed approach achieved overall true-positive rate of 99%, false-positive rate of 9%, precision of 91.7% and accuracy of 95.00%. Furthermore, we evaluated the effectiveness of our proposed psychological features. The results showed that psychological features can improve the accuracy of detection and reduce the false-positive rate. Our proposed method has a good performance in detecting phishing emails.

## 1. Introduction

Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumers' personal identity data and financial account credentials [1]. Phishers often try to use spoofed emails that claim to be from legitimate companies and institutions, misdirected recipients to click on a link that divulging personal information, or tricked recipients to return financial data such as password, account login credential or credit card numbers. According to APWG Phishing Attack Trends Reports [2], from April to June 2018, the number of new phishing emails exceeded 260,000, with an average monthly output of more than 80,000, and the number of detected new phishing websites increased to 230,000. Phishing activities are changing and growing at an alarming rate.

Many detection methods of phishing emails have been proposed in the literature to avoid the financial loss and information leakage caused by phishing activities. From an overview in existing work [3], despite the continuous improvement of the current phishing detection technology, it is still unable to effectively and accurately detect a wide variety of phishing attacks with high changing rate. It is of great necessity to select the features used in phishing emails to improve the adaptability and efficiency of detecting phishing emails. Moreover, we need to combine the research of social engineering in recent years to propose new features to improve the detection accuracy of phishing emails.

In this paper, we extract 18 features including header-based features, URL-based features, script-based 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 that

phishing 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.

## 2. 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 third-party 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.

## 3. Proposed Method

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.

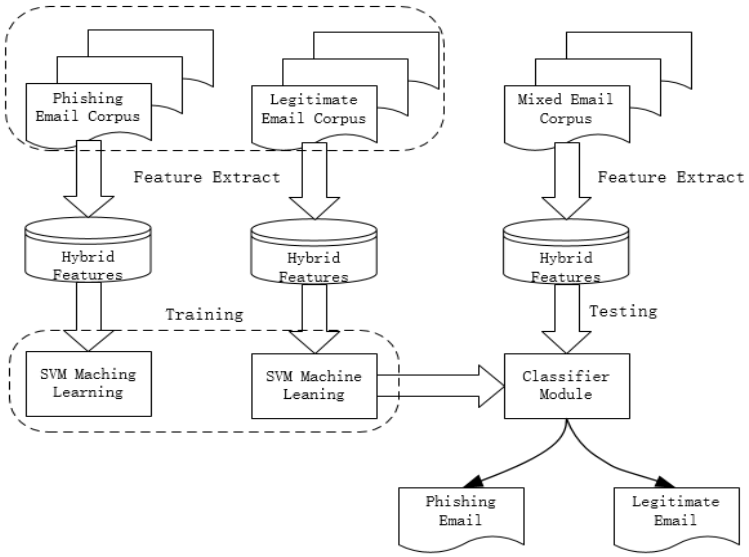


**Figure 1.** Architecture of Proposed Method

*3.1. 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 $\vec{x}_i$, and using $y_i$ to represent the label of i-th mail. If the mail is legitimate mail, $y_i$ equals 1. If the mail is phishing mail, $y_i$ is -1. Then the total feature set of n emails is:

$$D=(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots (\vec{x}_n, y_n) \quad (1)$$

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.

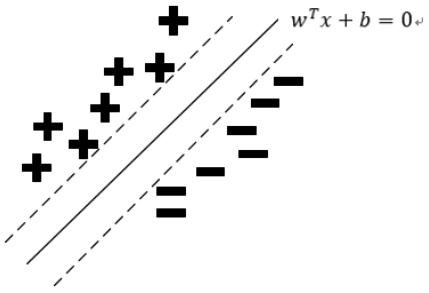


**Figure 2.** Support Vector Machine

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 $\vec{x}_i$ 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 $w^T$ 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.

### *3.2. 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 $x_i$ be the i-th feature of a mail, so the n-th mail's feature vector can be represented as $\vec{x}_n = (x_1 x_2 \ldots x_{18})$. The extracted features in this paper is shown in table Ⅰ.

**Table 1.** Feature List

| Feature | Description |
|---|---|
| $x_1$ | Blacklist words in title, such as bank, pay, account, password |
| $x_2$ | Inconsistency between replies address and send's address |
| $x_3$ | Inconsistency between sender domain and Message-Id domain |
| $x_4$ | Whether email is html format |
| $x_5$ | Whether presence of IP address in URL |
| $x_6$ | Whether presence of image links |
| $x_7$ | Whether presence of abnormal ports in URL |
| $x_8$ | The number of "%" character in URL |
| $x_9$ | The number of "@" symbol in URL |
| $x_{10}$ | The number of "." in URL |
| $x_{11}$ | The number of URLs in email |
| $x_{12}$ | The number of domains |
| $x_{13}$ | Inconsistency between where the URL actually leads to and the text the link appeared |
| $x_{14}$ | Whether presence of JavaScript in email |
| $x_{15}$ | Whether the JavaScript can change the status bar |
| $x_{16}$ | Whether the JavaScript can change the pop-up event |
| $x_{17}$ | Whether the JavaScript can change the on-Click event |
| $x_{18}$ | The psychological features in email |

*3.2.1. 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 $x_1$ 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 $x_2$ 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 $\boldsymbol{x}_3$ 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 $\boldsymbol{x}_4$ will be 1, otherwise, it will be 0.

*3.2.2. Header-based Features:* Phishing emails usually include URL to phishing websites in order to misdirect users to counterfeit websites that trick recipients to return personal data. Phishers have to use some means to hide the phishing links. In this paper, URL-based features consist the following nine features.

(1) Presence of IP address

In order to save the cost of domain name or prevent the identification of the real domain name, some phishers will directly use the host IP address as the URL. Therefore, IP address is the simplest feature to detect phishing links in emails. For example, if the URL of a mail is http://10.168.123.123/login, we can regard this email as a potential phishing email. Hence, if there are IP addresses in links, the feature value $\boldsymbol{x}_5$ will be 1, otherwise, it will be 0.

(2) Presence of image Links

The phishers will make the phishing emails look very complicated to increase the possibility that the recipient clicks on the phishing URL. One way to make the phishing email look complicated is to add lots of images to mail. These images will attach some phishing links. Hence, we use the presence of image links to detect phishing emails. If there are image links, the feature value $\boldsymbol{x}_6$ will be 1, otherwise, it will be 0.

(3) Abnormal port

Normal website will use '80' standard port commonly. Thus, if the port in the attached URL is not a standard port, such as '12345', '34232', then the email is a potential phishing email. If there are abnormal ports in URL, the feature value $\boldsymbol{x}_7$ will be 1, otherwise, is will be 0.

(4) Number of "%" character in URL

I After the attacker hexadecimal code the phishing URL, the URL still leads to the phishing website. But it is difficult for the recipient to directly identify the real address. For example, http://www.normal.com%2e%70%68%69%73%68%6d will redirect to a phishing website. Hexadecimal encoding will generate a large amount of "%" characters in URL. Therefore, we can make use of the number of "%" in URL to detect phishing emails. We set the feature value $\boldsymbol{x}_8$ be the number of "%" character.

(1) Number of "@" symbol in URL

If the URL consists "@" symbol, the browser will not lead to the address preceding the "@" symbol but lead to the real address following the "@" symbol. For example, the URL http://www.normal.com@phishing.com will not lead to normal.com but to phishing.com site. We count the number of links with the symbol, then set the feature value $\boldsymbol{x}_9$ be the number.

(2) Number of "." in URL

Normally, the URL of a legitimate email will not use subdomains. On the contrary, the attackers attempt to construct legitimate looking URLs by using subdomains. Hence, the number of "." in phishing URLs is greater than legitimate URLs. For example, http://apple.user_update.com will not lead to Apple websites instead to a phishing site. Hence, we count the number of dots, and set the feature value $\boldsymbol{x}_{10}$ to the number of dots.

(3) Number of URLs and domains

Usually there is only one domain name in a legitimate official email. In order to confuse recipients, phishers will make phishing emails contain legitimate URLs and phishing URLs. This makes the number of domain names and URLs in the mail more than one. Hence, counting the number of URLs and domains in an email is significant features. We set the feature value $\boldsymbol{x}_{11}$ be the number of URLs, then set the feature value $\boldsymbol{x}_{12}$ be the number of domains.

(4) Inconsistent URL

It is common for phishers to trick an unsuspecting recipient by offering a legitimate looking URL as a clickable button to a hyperlink, but in fact the URL will lead the recipients to a phishing site when is clicked. We will measure the inconsistency between the text the URL showed in the email and where the URL actually directs to. If we find such inconsistent URLs, let the feature value $\boldsymbol{x}_{13}$ be 1, otherwise, it will be 0.

*3.2.3. Script-based Features:* JavaScript in an email can easily steal user information, such as Cookie, password and so on. Legitimate emails will not contain well-designed JavaScript. This section details four features related to Script in emails.

(1) Presence of JavaScript

It is not uncommon for phishers to hide information from recipients by JavaScript. If we find JavaScript in emails, the emails are the most potential to be phishing emails. Then the feature value $\boldsymbol{x}_{14}$ will be 1, otherwise, it will be 0.

(2) Change the status bar

Phishers can deceive user by making the status bar of the browser not show the real address of phishing URL. The JavaScript is able to change the status bar easily. We can find whether JavaScript has such function. If we find it, the feature value $\boldsymbol{x}_{15}$ will be 1, otherwise, it is 0.

(3) Change Pop-up

We find it that many phishing emails will use JavaScript to change the pop-up event in order to trick the recipients to trust that this is a legitimate business website. If the email has the JavaScript to change the pop-up event, we set the feature value $\boldsymbol{x}_{16}$ be 1.

(4) Change on-Click

Many phishers also use JavaScript to change the on-Click event to make users not identify the phishing website. If the email has the JavaScript to change the on-Click event, we set the feature value $\boldsymbol{x}_{17}$ be 1.

*3.2.4. Psychologist Features:* Current research shows that people's psychological state is affected by the words they read. As a social engineering attack, phishing emails mainly deceive recipients by using the recipient's psychological weakness, instinctive reaction, curiosity, trust, anxiety and other psychological traps. Phishing emails often make the recipients nervous, fear, anxiety and worry to break through the recipient's psychological defense. For example, phishing email often informs the recipient of a problem with their account to get the account information. The content of such an email always contain negative words, such as fearful, hardly, nervous, urgent and so on, to indicate that if the recipient does not take action, there will be bad consequences. Based on Saif [10] works, we proposed nine types psychological bags of words including a total number of 108 keywords to classify the phishing emails. The psychologist words list is shown in Table 3.

The psychological features are calculated as follows: We calculate the proportion of each of the nine types of psychological words in the total number of words in the mail. For example, we denote the number of i-th psychological feature words appearing in the mail text as $C_i$, and use A to represent the

total number of words in the mail text. Then, the value of the i-th psychological feature $P_i$ of the mail is:

$$P_i = \frac{C_i}{A} \tag{4}$$

We set the feature value $\boldsymbol{x}_{18} = (P_1, P_2, \ldots P_9)$.

**Table 3.** Psychological Keywords

| Category | Keywords |
|---|---|
| negative | hardly, never, nothing, no, scarcely, |
| anxiety | anxious, nervous, worried, fearful |
| indignation | mad, annoyed, indignant, furious, blue, |
| sadness | sad, maze, guilty, error, mistake, |
| comprehension | understand, consider, aware, realize |
| hesitation | maybe, perhaps, hesitate, indecisive |
| certainty | always, indeed, sure, affirmative |
| repression | constrain, stop, block, desperate |
| trust | faithful, fortune, loving, kind, promise |

## 4. Experiment

The experimental datasets are collected from two publicly available datasets which contains phishing emails and legitimate emails. The phishing email dataset is from https://monkey.org/~jose/phishing/. The legitimate emails are collected from CSDMC2010. From these datasets, we randomly selected 500 phishing emails represented as $Phishing$ and 500 legitimate emails represented as $Legit$. The experiments selected 60 percent of the dataset for training dataset and the remaining 40 percent for testing dataset.

### *4.1. Evaluation Metrics*

In order to demonstrate the performance of our proposed method, this work chose four commonly used evaluation metrics, which are Precision, Accuracy, True Positive Rate (TPR) and False Positive Rate (FPR) [11]. TPR, FPR, Precision and Accuracy are defined as follows:

$$\text{TPR} = \frac{Phish_{phish}}{Phish_{phish} + Phish_{legit}}$$

$$\text{FPR} = \frac{Legit_{phish}}{Legit_{phish} + Legit_{legit}}$$

$$\text{Precision} = \frac{Phish_{phish}}{Phish_{phish} + Legit_{phish}}$$

$$\text{Accuracy} = \frac{Phish_{phish} + Legit_{legit}}{Phishing + Legit}$$

$Phish_{phish}$ represents to the number of phishing emails that were classified to be phishing emails, $Phish_{legit}$ represents to the number of phishing emails that were classified to be legitimate emails, $Legit_{legit}$ represents to the number of legitimate emails classified to be legitimate emails, and $Legit_{phish}$ represents to the number of legitimate emails classified to be phishing emails. $Phishing$ represents to the total number of phishing emails. $Legit$ represents the total number of legitimate emails.

*4.2. 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 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 | $\mu_{phish}$ | $\sigma_{phish}$ | $\mu_{legit}$ | $\sigma_{legit}$ |
|---|---|---|---|---|
| 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 |

**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.

## 5. 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] Naghmeh Moradpoor, Benjamin Clavie and Bill Buchanan. “Employing machine learning techniques for detection and classification of phishing emails,” Computing Conference, 2017.
[9] May, Lew et al. Phishing Email Detection Technique by using Hybrid Features, 9th International Conference on IT in Asia (CITA), 2015
[10] Saif M. Mohammad. Sentiment “Analysis of Mail and Books”. Technical report, National Research Council Canada, 2011.
[11] Adewumi, Oluyinka Aderemi and Akinyelu, Ayobami Andronicus, “A hybrid firefly and support vector machine classifier for phishing email detection,” Kybernetes, 2016, vol. 45, no. 6, pp. 977–994.