*International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868*
*Foundation of Computer Science FCS, New York, USA*
*Volume 12 – No. 7, October 2017 – www.ijais.org*

# Phishing Detection in E-mails using Machine Learning

Srishti Rawal
VIT University
Chennai, India

Bhuvan Rawal
BITS Pilani, Goa,
India

Aakhila Shaheen
BITS Pilani
Hyderabad, India

Shubham Malik
VIT University
Chennai, India

## ABSTRACT
Emails are widely used as a means of communication for personal and professional use. The information exchanged over mails is often sensitive and confidential such as banking information, credit reports, login details etc. This makes them valuable to cyber criminals who can use the information for malicious purposes. Phishing is a strategy used by fraudsters to obtain sensitive information from people by pretending to be from recognized sources. In a phished email, the sender can convince you to provide personal information under false pretenses. This experimentation considers the detection of a phished email as a classification problem and this paper describes the use of machine learning algorithms to classify emails as phished or ham. Maximum accuracy of 99. 87% is achieved in classification of emails using SVM and Random Forest classifier.

## General Terms
Phishing, security, classification

## Keywords
Phishing detection, SVM, ham, naive bayes, machine learning, email fraud, artificial intelligence

## 1. INTRODUCTION
Phishing is a lucrative type of fraud in which the criminal deceives receivers and obtains confidential information from them under false pretenses. Phished emails may direct the users to click on a link of a website or attachment where they are required to provide confidential information like passwords, credit card information etc. The phisher sends out the messages to thousands of users and usually only a small percentage of recipients may fall into the trap but this can result in high profits for the sender.

In 2006, hackers in America used emails as a mode of setting "baits" for users to steal usernames and passwords of American Online accounts. Ever since then the techniques of phishing have evolved making it harder to identify fraudulent emails. As per the 2016 data breach report by Verizon, roughly 636,000 phishing emails were sent out of which only 3% of the targeted individuals alerted the management of a possible phished emails.

A massive phishing attack targeting millions of Gmail users hit google in May 2017, in which the hacker gained access email histories of users. Through this information, the hackers were able to pose emails as belonging to a known source and asked them to check the attached file. On clicking the link to attacked file, the users were asked to give permission for a fake app to manage users email account.

With the ever increasing use of emails and growth of technologies, risk of losing valuable information to fraudsters has also been increasing. This paper focuses on identifying a phished email with the help of machine learning algorithms.

In the proposed system, detecting phished email can be described as a classification problem with two categories i.e. ham and phished. Machine Learning is a field of artificial intelligence in which the system is given the ability to learn without being explicitly programmed. In our model, supervised machine learning algorithms are used for classification. Supervised learning algorithms predict the nature of unknown data based on the known examples. These algorithms are a subset of machine learning algorithms which iteratively learn from data.

The remainder of the paper is organized as follows. Section 2 discusses the existing systems used for detection of phishing in emails. The third section describes proposed system, the algorithms used and provides a brief description of the features used. Further, in section 4, the results obtained are explained. In the fifth section, a conclusion is drawn and followed by this is the reference section.

## 2. RELATED WORK
Andronicus et al. used random forest machine learning classifier is used for classification of phished emails. They have aimed to maximize the accuracy and minimize the number of features required for classification. A content-based phishing detection approach which has high accuracy is presented.

In [2], authors proposed a model based on extracted features which appear in the header and HTML body of email which are classified using feed forward neural network. The results indicate 98.72% accuracy of classification.

In [3], over 7000 emails are used in dataset and a number of different features used. Overall accuracy of 99.5% is achieved.

Gilchan Park et. al. aimed to extract robust features in order to discriminate legitimate and phished emails. A comparison of sentence syntactic similarity and the difference in subjects and objects of target verbs between phishing emails and legitimate emails is done.

In "Email Phishing : An open threat to everyone", the different techniques of phishing are analyzed and suggestions for users to avoid falling into the trap of fraudsters are provided.

C. Emilin Shyni et al. proposes a methodology incorporating natural language processing, machine learning and image processing is described. They use a total of 61 features are used. They achieved an classification accuracy of above 96% using a multi-classifier.

In "Detection Phishing Emails Using Features Decisive values", 18 features are extracted and the proposed algorithm classifies each email depending upon existence of flags and weightage of features. Their results show that out of the 18 features extracted, high accuracy of can be obtained if most effective features are used for classification

In "Phish-IDetectore" authors focus on the properties of Message-IDs and apply n-gram analysis to the Message-IDs.

They applied different machine learning techniques for classification and claim detection rates of above 99%.

# 3. PROPOSED SYSTEM

For the purpose of classification, 9 features were extracted from all emails in a self-made dataset which consists of n number of phished emails and m number of ham emails . These features are fed into the classifiers and results noted. Aim is to use the least number of features to develop a system which provides higher accuracy and study the variation of features.

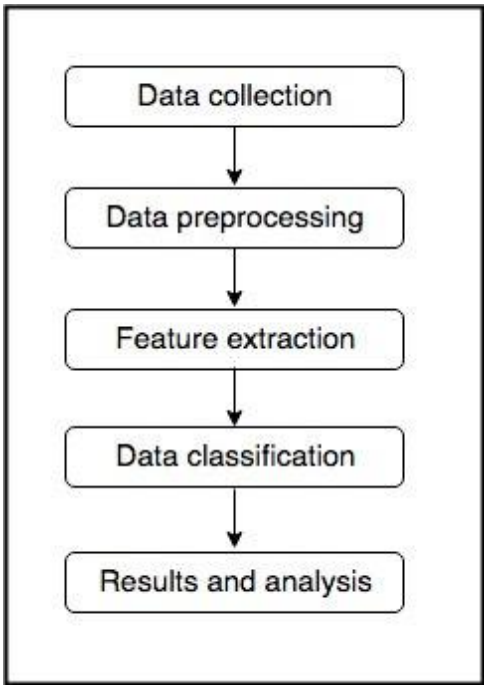


**Figure 1: Process flow diagram**

## 3.1 Features

This section will describe the features extracted.

### *3.1.1 Link based*

Domain count: In order to make the links look legitimate, attackers add subdomains to the links. Adding subdomains, increased the number of dots in the link. As proposed by Emigh, the number dots in a legitimate email should not be more than 3 [3]. This is a binary feature i.e. if there exists a link in the mail which has number of dots greater than 3, it would be considered a phished email.

Number of links: Phished emails generally contain greater number of links as compared to ham since the sender aims to redirect the user to an illegitimate website by deceiving him. This is a continuous feature

### *3.1.2 Tag based*

Presence of javascript: Presence of javascript in an email suggests that the sender is either trying to hide information or activate certain changes in the browser [9]. This is a binary feature. If the <script> tag is present in the mail then it is considered to be a phished mail.

Presence of form tag: In order to obtain information from users phished emails contains forms embedded in them. This is a binary feature i.e. presence of form tag indicates that it is a phished email.

Presence of HTML: HTML emails enable the sender to include embedded images and hyperlinks in the mail which plan text emails do not support. If html tag is present in the email, it is considered to be phished. This is a binary feature.

### *3.1.3 Word based*

Number of action words: Presence of action words in emails indicates whether the sender is expecting a response from the user to perform certain action such as clicking on a link, filling a form, providing certain information etc. This is a continuous feature.

Presence of word paypal: Often, the sender pretends to be a part of organizations which seem legitimate. Presence of the word paypal in the links of the mail or in the “from” section would suggest that the sender is associated with paypal. This is a binary feature.

Presence of word bank: This is a binary feature suggesting that the mail is related to banking information. The sender would either be pretending to be a member of the banking organization or seeing the reader’s credentials.

Presence of word account: This would suggest that the email is looking for email related to an account. It can be a social media account or bank account etc. It is a binary feature.

Combining the three types of features described in 3.1.1, 3.1.2 and 3.1.3, a total of 9 different features are obtained which are extracted with the help regular expressions and Python’s NLTK (natural language toolkit).

## 3.2 Classifiers

This section will give a detailed description of the classifiers used

### *3.2.1 Support Vector Machines*

SVM is a supervised algorithm which is popular for text classification algorithm due to high speed and good performance. Based on the training set provided, it outputs a hyperplane which is a line in two dimension that best separates the categories. This hyperplane is called the decision boundary. In phishing detection, input is represented by a set of features for instance, presence or absence of certain word and output which is 1 or -1 indicates whether the email is phished or not.

### *3.2.2 Naive Bayes*

The naive bayes classifier belongs to the family of probabilistic algorithms and used bayes theorem to categorize sample data.

Bayes theorem : Given a hypothesis H and evidence E, Bayes' theorem states that the relationship between the probability of the hypothesis P(H) before getting the evidence and the probability P(H|E) of the hypothesis after getting the evidence is :

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$$

The probability of each category is calculated and outputs the one with highest probability.

### *3.2.3 Random Forest*

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set.

### 3.2.4 *Logistic regression*

The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the probability of a given outcome by a specific percentage.

### 3.2.5 *Voted Perceptron*

This algorithm stores all weight vectors and let them vote on test examples. It is fast, simple and has been claimed to be as good as support vector machines in many situations.

## 3.3 Dataset

The dataset used comprises of 1605 emails out of which 1191 are ham and 414 are phished. The ham emails are collected from a publicly available dataset and phished emails are a combination of emails from various sources.

# 4. RESULTS AND DISCUSSION

The dataset consisting of extracted features is partitioned then fed into five classifiers and results noted.10 fold cross validation technique has been used for partitioning the original data sample into a training set and test set.

**K-fold cross validation**: In k fold cross validation, the dataset is randomly split into k mutually exclusive subsets of approximately equal sizes [10] . Followed by this, the model is trained and tested k times, of the k samples, a single subsample is retained as validation data of the testing model and remaining k-1 subsamples are used as training set.

It is observed that tree based, SVM and logistic classifiers classify most accurately. Performance of different classifiers is evaluated using different performance metric which are described in this section. It is observed that SVM and Random Forest classify the dataset with highest accuracy of 99.87%. The following performance metrics are used for evaluating our model:

**Precision**: It is defined as the fraction of retrieved objects that are relevant [9]. In our case it is the fraction of emails that are correctly classified as phished which are actually phished.

$$Precision = \frac{TP}{TP + FP}$$

**Recall**: It is defined as the proportion of relevant objects that are retrieved relative to the total number of relevant objects in the dataset [9] i.e. the fraction of phished emails which are classified are phished from the dataset.

$$Recall = \frac{TP}{TP + FN}$$

**F-measure**: It is defined as the harmonic mean of precision and recall [8].

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

**True Positive Rate**: The percentage of phished emails in the dataset that are correctly classified as phished. Let P be the number of phished emails and $N_p$ be the number of correctly classified phished emails then true positive rate can be calculated as:

$$TP = \frac{N_p}{P}$$

**True Negative Rate**: The percentage of ham emails in the dataset correctly classified as ham. Let H be the number of ham emails and $N_h$ be the number of correctly classified ham emails then true negative rate can be calculated as:

$$TN = \frac{N_h}{H}$$

**False Positive Rate**: The percentage of ham mails incorrectly classified by the model as phished. Let $N_f$ be the number of ham emails incorrectly classified as phished and number of ham emails is H then false positive rate can be calculated as:

$$FP = \frac{N_f}{H}$$

**False Negative Rate**: The percentage of phished emails that were incorrectly classified by the model as ham. Let $P_h$ be the number of phished emails that are classified as ham and P be the number of phished emails then false negative rate can be calculated as:

$$FN = \frac{P_H}{P}$$

**Table 1: Comparison of Precision, Recall, F- measure (weighted average)**

| Classifier | Precision | Recall | F-measure |
|---|---|---|---|
| SVM | 0.999 | 0.999 | 0.999 |
| Random Forest | 0.999 | 0.999 | 0.999 |
| Logistic | 0.999 | 0.999 | 0.999 |
| NaiveBayes | 0.998 | 0.998 | 0.998 |
| VotedPerceptron | 0.956 | 0.956 | 0.956 |

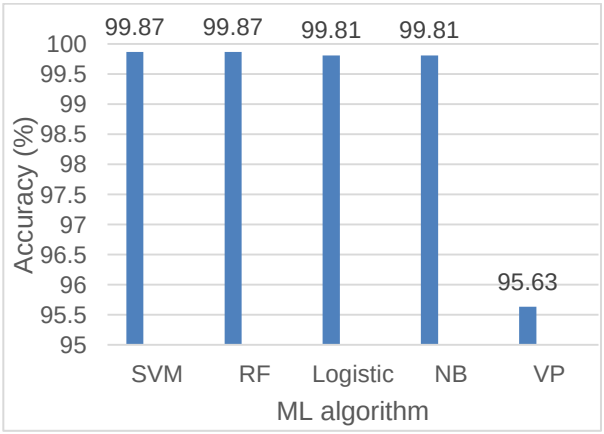


**Figure 2: Classification Accuracy of 5 ML Classifiers**

**Table 2: Comparison of Accuracy**

| Classifier | Accuracy (%) |
|---|---|
| SVM | 99.87 |
| Random Forest | 99.87 |
| Logistic | 99.81 |
| NaiveBayes | 99.81 |
| VotedPerceptron | 95.63 |

**Table 3: Comparison of true positive and true negative (weighted average)**

| Classifier | TP | FP |
|---|---|---|
| SVM | 0.999 | 0.002 |
| Random Forest | 0.999 | 0.002 |
| Logistic | 0.998 | 0.002 |
| NaiveBayes | 0.998 | 0.002 |
| VotedPerceptron | 0.956 | 0.083 |

From these results, it is evident that SVM and Random Forest give better performance in terms of classification accuracy as compared to others. Table 1 shows the precision, recall and f-measure of the classifiers used. SVM, Random Forest and Logistic classifiers give 99.99% precision, recall and f-measure rates. Table 3 compares the true positive and true negative rates. It shows that SVM and Random Forest produce the highest true positive rates. Thus, it is clear that overall performances of SVM and Random Forest are better as compared to other classifiers in terms of accuracy, recall and precision.

The results show our model produces high accuracy in detecting phished emails. By using the most relevant features, the number of features has been reduced as compared to other works but at the same time, accuracy is improved.

## 5. CONCLUSION

This paper discusses an approach for classification of mails into phished and ham with the help of machine learning algorithms. The dataset was preprocessed and converted to a suitable form that could be fed into classifiers by extraction of relevant features. The features are extracted with the help of python programming language using regular expressions and NLTK. These are stored in a suitable file which is fed into different classifiers. Supervised learning algorithms have been used which require a training set using which they are able to categorize the test set. In order to partition the dataset, 10 fold cross validation technique has been used. The model is fed into SVM, Random Forest, Logistic, Naive Bayes and Voted Perceptron classifiers. The classification results were encouraging as the highest accuracy of 99.8% was achieved. This work has produced encouraging results, however, the dataset used may not necessarily replicate real life scenario . In future works, the proposed system can be improved by increasing the dataset. By adding a variety of emails both of type phished and ham, the system would be closer to the real life scenario where fraudsters are day by day improving their techniques. Using real life samples would enable us to deploy a formal system that can be used across organization and privately to prevent users from being victims to phishing attacks.

## 6. REFERENCES

[1] Verizon, Data Breach Report 2016

[2] Andronicus A. Akinyelu and Aderemi O. Adewumi. Classification of Phishing Email using Random forest Machine Learning Technique 2014.

[3] Noor Ghazi M. Jameel, Loay E. George. Detection of Phishing Emails using Feed Forward Neural Network, International Journal of Computer Applications 2013.

[4] Ian Fette, Norman Sadeh, Anthony Tomasi, Learning to Detect Phishing Emails, In Proceedings of the International World Wide Web Conference (WWW), 2006

[5] Gilchan Park, Julia M. Taylor, Using Syntactic Features for Phishing Detection 2015, https://arxiv.org/ftp/arxiv/papers/1506/1506.00037.pdf

[6] Gori Mohamed .J, M. Mohammed Mohideen, Mrs. Shahira Banu. Email Phishing - An open threat to everyone, International Journal of Scientific Research Publications, 2014

[7] C. Emilin Shyni, S. Sarju, S. Swaminathan A Multi-Classifier Based Prediction Model for Phishing Emails Detection Using Topic Modelling, Named Entity Recognition and Image Processing, SciRes 2016

[8] Noor Ghazi M. Jamee , Loay E. George (2014), "Detection Phishing Emails Using Features Decisive Values",257-259

[9] Rakesh M. Verma and Nirmala Rai. Phish-IDetector: Message-Id Based Automatic Phishing Detection, International Joint Conference on e-Business and Telecommunications 2015 .

[10] Basnet R., Mukkamala S., Sung A.H. (2008) Detection of Phishing Attacks: A Machine Learning Approach. In: Prasad B. (eds) Soft Computing Applications in Industry. Studies in Fuzziness and Soft Computing, vol 226. Springer, Berlin, Heidelberg

[11] Adwan Yasin and Adbelmunem, An intelligent classification model for phishing email detection , International Journal of Network Security & Its Applications (IJNSA) Vol.8, No.4, July 2016

[12] D. J. Hand, Heikki Mannila, Padhraic Smyth. Principles of Data Mining

[13] Ron Kohavi, A study of cross validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence, 1995.