Machine Learning Based Prediction versus Human-as-a-Security-Sensor

Safwana Haque[a,1*], George Loukas [a,2]

Department of Computing and Information Systems, Faculty of Architecture, Computing and Humanities, University of Greenwich, United Kingdom
[1] safwanahaque@gmail.com; [2] G.Loukas@greenwich.ac.uk
** corresponding author*

ARTICLE INFO

Article history:
Received : 23 Nov 2018
Revised : 23 Dec 2018
Accepted : 03 Feb 2019

Keywords:
Phishing
Feature extraction
Machine learning
Prediction
Classifiers
Logistic regression

ABSTRACT

Phishing is one of the most common cyber threats in the world today. It is a type of social engineering attack where the attacker lures unsuspecting victims into carrying out certain tasks mostly to steal personal and sensitive information. These stolen information are exploited to commit further crimes e.g. blackmails, data theft, financial theft, malware installation etc. This study was carried out to tackle this problem by designing an anti-phishing learning algorithm to detect phishing emails and also to study the accuracies of human phishing prediction to machine prediction. A graphical user interface was designed to emulate an email-client system that popped-up a warning on detecting a phishing mail successfully and collection of predictions made by expert and non-expert users on anti-phishing techniques. These predictions were compared to the predictions made by the machine learning algorithm to compare the efficiencies of all predictions considered in this research. The performance of the classifier used was measured with metrics such as confusion matrix, accuracy, receiver operating characteristic curve and area under graph.

I. Introduction

Phishing is a cyber attack that aims at gathering sensitive personal information such as financial account login credentials, credit card details, social security number, etc. through human interaction and persuasion and is thus known as a type of social engineering attack. The most common and widely used media of carrying out phishing attacks are via the internet using emails, web URLs (universal resource locator), instant messengers or web forums. In these instances, an attack could be implemented in various ways. For example, a fake URL may be provided that opens a phishing webpage that looks identical to the original webpage. This cloned webpage belongs to the phisher (attacker) and usually requests the user for login details or other sensitive credentials, which if provided go directly to the attacker. These pieces of information can then be used by the attacker for various malicious purposes such as transferring funds from bank accounts, taking out loans, taking over control of email accounts, etc. Other times, a link or attachment could be sent that downloads and installs a keylogger or a malware on the user's computer that stealthily collects information and transfers them to the attacker.

Other media of phishing attacks are via phone calls known as vishing while those done through text messages are known as smishing. Smishing is carried out by sending fake or malicious links embedded in SMSs (short messaging system) similar to normal phishing scams. In vishing, the scammer places a voice call to the victim impersonating a personnel from an authentic organization e.g. a bank or IT organisation, persuading the victim of an imminent problem that the scammer is willing to help in solving. Depending on the purpose of the particular vishing, the scammer could either convince the victim to divulge sensitive information or carry out certain tasks on the computer

DOI: *10.29099/ijair.v3i1.83*

that hand over total control to the attacker. The success of any of the phishing attacks above is highly dependent on the power of persuasion of the attacker.

Phishing attacks are further classified into (mass) phishing, clone phishing, spear phishing and whaling. The most common type of phishing exploits seen every day are of the mass phishing nature. They are not directed to anyone in particular but are usually sent in random and in bulk. Mass phishing attacks are usually successful and popular because the attackers do not need to perform any background research on the recipients unlike in spear phishing or whaling. Also, the attacks exploit the most commonly used services by people e.g. PayPal, Amazon, banks, etc. with the hope that the victim may be a subscriber of such a service. Clone phishing attacks make use of previously used or sent authentic mails from known companies or individuals that have been cloned but changed to contain malicious links or attachments. The sender's address is spoofed in any of these cloned mails to look as if the mail has actually been sent by the original sender thereby deceiving the receiver into believing that such a mail is authentic. Spear phishing is a phishing attack that is directed to an individual or a group of individuals. It appears to come from a trusted source to the victim e.g. from a fellow colleague, IT technician, etc. The targets of spear phishing are usually financial gains, trade secrets, data and identity thefts, etc. Whaling is a type of spear phishing attack where attacks are directed to personnel in powerful and topmost positions who are lured into opening emails and links that appear to be of great importance e.g. legal subpoenas, management issues and customer complaints. Successful whaling attacks usually incur the greatest losses.

Phishing is a cyber attack that aims at gathering sensitive personal information such as financial account login credentials, credit card details, social security number, etc. through human interaction and persuasion and is thus known as a type of social engineering attack. The most common and widely used media of carrying out phishing attacks are via the internet using emails, web URLs (universal resource locator), instant messengers or web forums. In these instances, an attack could be implemented in various ways. For example, a fake URL may be provided that opens a phishing webpage that looks identical to the original webpage. This cloned webpage belongs to the phisher (attacker) and usually requests the user for login details or other sensitive credentials, which if provided go directly to the attacker. These pieces of information can then be used by the attacker for various malicious purposes such as transferring funds from bank accounts, taking out loans, taking over control of email accounts, etc. Other times, a link or attachment could be sent that downloads and installs a keylogger or a malware on the user's computer that stealthily collects information and transfers them to the attacker.

Other media of phishing attacks are via phone calls known as vishing while those done through text messages are known as smishing. Smishing is carried out by sending fake or malicious links embedded in SMSs (short messaging system) similar to normal phishing scams. In vishing, the scammer places a voice call to the victim impersonating a personnel from an authentic organization e.g. a bank or IT organisation, persuading the victim of an imminent problem that the scammer is willing to help in solving. Depending on the purpose of the particular vishing, the scammer could either convince the victim to divulge sensitive information or carry out certain tasks on the computer that hand over total control to the attacker. The success of any of the phishing attacks above is highly dependent on the power of persuasion of the attacker.

Phishing attacks are further classified into (mass) phishing, clone phishing, spear phishing and whaling. The most common type of phishing exploits seen every day are of the mass phishing nature. They are not directed to anyone in particular but are usually sent in random and in bulk. Mass phishing attacks are usually successful and popular because the attackers do not need to perform any background research on the recipients unlike in spear phishing or whaling. Also, the attacks exploit the most commonly used services by people e.g. PayPal, Amazon, banks, etc. with the hope that the victim may be a subscriber of such a service. Clone phishing attacks make use of previously used or sent authentic mails from known companies or individuals that have been cloned but changed to contain malicious links or attachments. The sender's address is spoofed in any of these cloned mails to look as if the mail has actually been sent by the original sender thereby deceiving the receiver into believing that such a mail is authentic. Spear phishing is a phishing attack that is directed to an individual or a group of individuals. It appears to come from a trusted source to the victim e.g. from a fellow colleague, IT technician, etc. The targets of spear phishing are usually financial gains, trade secrets, data and identity thefts, etc. Whaling is a type of spear phishing attack where attacks are directed to personnel in powerful and topmost positions who are lured into

opening emails and links that appear to be of great importance e.g. legal subpoenas, management issues and customer complaints. Successful whaling attacks usually incur the greatest losses.

II. Research Method

To achieve the objectives stated earlier, some essential tools and techniques were required such as an appropriate machine learning algorithm for prediction, a programming language to design the machine learning algorithm, a programming language to design the email-client GUI and a database to store the various phishing and ham mails. These tools and techniques are described below.

A. Machine learning Process

Machine learning is the process of extracting knowledge from data and making predictions with the use of an algorithm or model in either a supervised or an unsupervised manner. In supervised machine learning, a set of inputs is given with their corresponding outputs from which the algorithm or model tries to predict future results. In an unsupervised learning on the other hand, inputs are given with no corresponding outputs and the algorithm tries to figure out a pattern to predict future behaviours. Machine learning problems are broadly categorised into classification, regression and clustering depending on the outcome of the desired output or result. In classification problems, the inputs are labelled into at least two or more discrete classes such that the algorithm has to predict results into one of these groups appropriately. Regression problems are different from classification in that the classes are continuous rather than discrete while in clustering problems, the inputs fall into uncategorised or unknown groups or clusters.

Since one of the objectives of this project was to predict if a mail were a phishing or a ham, a supervised classification machine learning algorithm was required. A logistic regression machine learning algorithm was selected to achieve this as it is a linear statistical model that is commonly used to implement classification problems with binary outputs. It was implemented in this study using Python, one of the programming languages famously used for implementing machine learning algorithms. The distribution of python used for this project was Anaconda and it was used along with scikit-learn, which provides a detailed and well-documented machine learning library. After outlining the problem to solve and objectives to achieve, selecting the tools to use and setting up the environment to carry out the machine learning segment of this project, four other major steps were taken as shown in Figure 1.

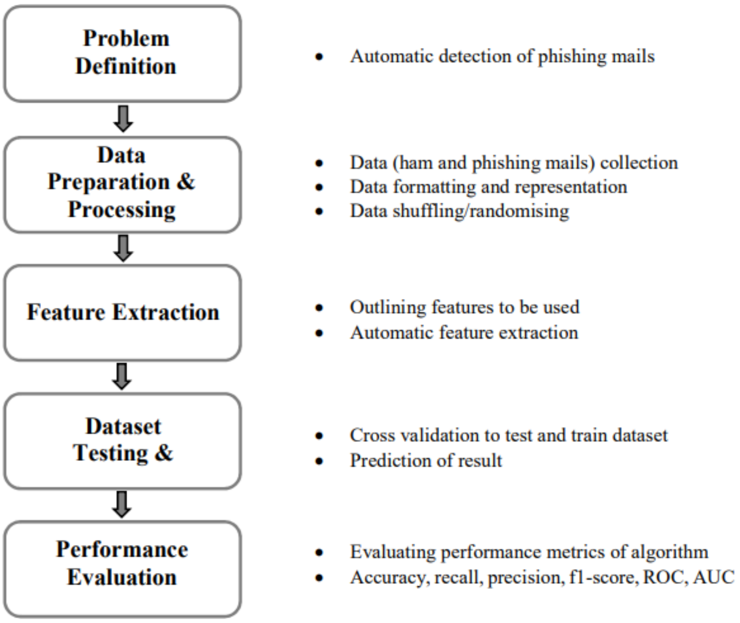

Problem Definition
Automatic detection of phishing mails
Data Preparation & Processing
Data (ham and phishing mails) collection
Data formatting and representation
Data shuffling/randomising
Feature Extraction
Outlining features to be used
Automatic feature extraction
Dataset Testing &
Cross validation to test and train dataset
Prediction of result
Performance Evaluation
Evaluating performance metrics of algorithm
Accuracy, recall, precision, f1-score, ROC, AUC

Fig. 1. Steps carried out in the machine learning process

In the data preparation and processing stage, both ham and phishing mails were collected from the internet and personal inboxes. For each mail, the essential parts (the sender's email address, the receiver's email address, the subject of the mail and body of the mail) needed for the feature extraction in the next stage were stored in columns in a text file. A binary tag (0/1) was given to each mail to indicate if the mail were a ham or a phishing mail. These mails were then randomly shuffled so that there was no fixed pattern of storing or selecting mails. The feature extraction was done by the JAVA segment of this project from the 100 emails collected above. Fourteen features that are typically characteristics of phishing mails were automatically extracted and stored in a comma separated value (CSV) file. A '1' was used to indicate the presence of a feature while a '0' indicated an absence of the feature. These features were collected to be used by the machine learning algorithm to predict a mail as ham or phishing and are briefly explained below.

- Use of generic salutation: checks to see if a mail starts with generic salutations like 'Dear User', 'Dear Subscriber', 'Dear Member' and 'Dear Customer' or with simply a 'Hi/Hello'. Mails that did not start with any salutation were also considered to start with a generic salutation.
- Address mismatch: mails that had different sender and reply-to addresses were considered to have this feature.
- Presence of a link: program checks to see if a mail contains one or more links in the mail.
- Link length: a link length longer than 25 characters was flagged. Links that contained IP addresses were flagged as these are usually used to camouflage harmful links.
- Dots: links having more than three dots were considered suspicious and flagged with a value of '1'.
- Domains: mails having multiple domains were also flagged e.g. www.google.com.my_goo.gle.co.uk/login_page.jsp is a URL with multiple domain names.
- Click URLs: links like 'Click here', 'Click below', 'Click the following URL/link' found in a mail were indicated with a '1' in this column.
- Keywords A: the program searched for words like 'update', 'confirm', 'upgrade' and, if found, this column was stored with a '1'.
- Keywords B: words like 'suspend', 'restrict', 'hold', 'block', 'lose', 'cancel', 'expire' were searched in the body of the mail. If any or more were found, this mail was flagged and a '1' was used to indicate so.
- Keywords C: mails with words like 'verify' and 'account' were flagged.
- Keywords D: in this group, words like 'login', 'username', 'password' were considered suspicious.
- Keywords E: mails requesting information like 'SSN' (social security number), 'bank account number', 'pin number' and 'credit card number' were considered suspicious.
- Urgent status: mails that portrayed a sense of urgency with words like 'urgent', 'important', 'notification', '24 hours', '48 hours', 'immediately' and '14 days' were flagged as suspicious.

A sample of the extracted features is shown in Figure 4. After the identification and extraction of a phishing mail's features, comes the testing and training stage of the machine language classifier. In this stage, part of the mails collected was used to train the logistic regression classifier and the other part was used to test the accuracy and efficiency of the classifier. The basic and simple method used for testing and training a classifier is to divide the data into either a 70:30, 75:25 or 80:20 test/train ratio.

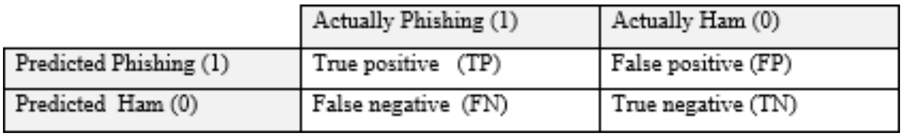
However, a more sophisticated test/train method was used in this project known as cross-validation. In cross-validation, the entire dataset is divided into a number of groups called folds where one of the folds is used as the test set while the rest of the folds are used as the train sets. This test/train process is iterated for the number of folds that the dataset is divided into. This type of cross-validation is called k-fold cross-validation and for this study, a 10-fold cross-validation was used to test and train this machine learning algorithm. Hence, the dataset was divided into 10-folds where each fold contained 10 unique mails. The iteration process was carried out 10 times, changing the test and train sets each time. The predictive result is the average result of all the individual results gotten after every iteration.

The accuracy of cross-validation test/train method is usually higher and better than simple test and train methods and that is why this method was opted for.

The research is aimed towards designing and implementing a plant recognition system. It is important to recognize plants as it can be used for medicinal, educational or herbal purposes. The research plan is an important part as it helps in gathering and answering important questions that are related to the research. The steps taken are as described below.

After the machine learning model was tested and trained to predict results, its performance was evaluated to understand its degree of accuracy and efficiency. The most common performance metrics used and considered for this project were accuracy, precision, recall, f1-score, ROC and area under curve (AUC). However, to calculate these metrics, the concepts and values of true positives/negatives and false positives/ negatives have to be considered.

True positives (TP) is the number of data outputs or labels that were correctly identified as positives (i.e. as phishing mails) by the machine learning algorithm while true negatives (FP) is the number of data outputs or labels that were correctly identified as negatives (i.e. as ham mails). False positives (FP) are instances when the classifier identified negative data outputs (i.e. ham mails) as positives (i.e. as phishing mails) while false negatives (FN) are instances when the classifier identified positive data outputs (i.e. phishing mails) as negatives (i.e. ham mails). For easier understanding, this is depicted using a confusion matrix in Figure 2.



	Actually Phishing (1)	Actually Ham (0)
Predicted Phishing (1)	True positive (TP)	False positive (FP)
Predicted Ham (0)	False negative (FN)	True negative (TN)

Fig. 2.Confusion matrix of phishing and ham mails

Brief explanations of the performance evaluation metrics used in this project and their formulae based on the confusion matrix are given below.

- Accuracy: is the rate of correct predictions that the machine learning classifier makes from its overall predictions, which is calculated using (TP+TN)/(TP + FP+FN+TN)

- Specificity or true negative rate (TNR): is the measure of the correctly identified true negatives by the classifier from all possible negatives and it is calculated by TN/(TN + FP)

- Precision (P) or positive predictive value (PPV): is a measure of number of times that the true positives predicted is right, calculated using TP/(TP+FP)

- Recall (R), sensitivity or true positive rate (TPR): is the measure of the correctly identified true positives by the classifier from all possible positives and is calculated using TP/(TP+FN)

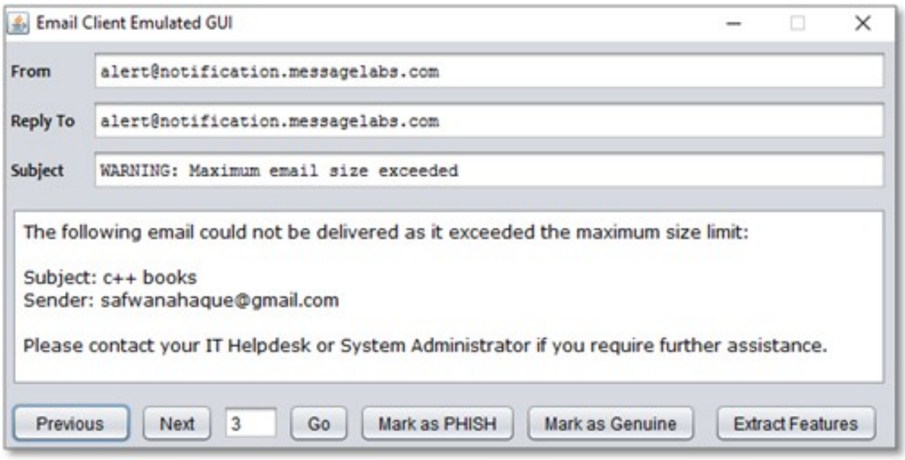
- F1-score, f-score or f-measure is a measure of the accuracy of the test and is also known as the harmonic mean of precision and recall. It is calculated using 2(P*R)/(P+R)

- ROC (receiver operating characteristic) curve is a statistical graph that is achieved by plotting TPR against FPR (false positive rate) also known as fall-out or false alarm. FPR is a measure of incorrectly predicted positives from all possible negatives (or 1 specificity) as shown: FP/(TN+FP) AUC: goes hand-in-hand with ROC and is computed by calculating the area under the ROC curve. It is a numerical measure of the overall performance of the classifier. A value above 0.5 is considered better than guessing while anything less than that is worse.

The results of the evaluation of the classifier using these metrics are discussed in section 4. A GUI was used to act as an interface between the system and a user. It was designed to emulate a simplified real life email-client interface with custom features such as ‘previous’, ‘next’, ‘mark as phish’, ‘mark as genuine’ and ‘extract features’. It was designed with the use of JAVA, a high level programming language, which was selected because of a number of reasons such as its ease of use, its cross-platform compatibility, high performance, rich tool palette for design, good user support and documentation.

An instance of the GUI is shown in Figure 3, which displays one of the 100 mails used in this study. The ‘from’ and ’reply-to’ textboxes show the email addresses of the sender and receiver of the current mail in view respectively, the ‘subject’ textbox shows the subject of the mail, the text area displays the content of the mail, the ‘previous’ and ‘next’ buttons are for navigating between mails, the textbox at the bottom of the window shows the number of the mail shown, the ‘go’ button is used to navigate to a different mail when a number is put in the previous textbox, the ‘mark as phish’ and ‘mark as genuine’ buttons were used to mark the current mail as either phish or ham. These buttons were provided to collect feedbacks from both expert and non-expert users on phishing for all 100 mails.

These feedbacks were stored in a CSV file for analysis purpose which are further discussed in section



From: alert@notification.messagelabs.com

Reply To: alert@notification.messagelabs.com

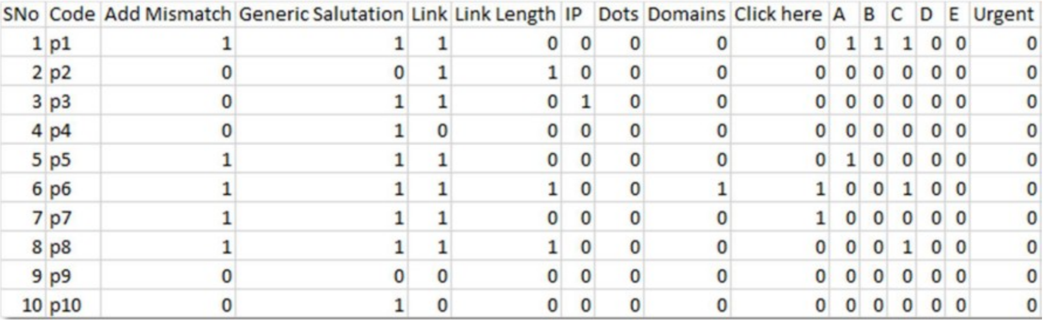
Subject: WARNING: Maximum email size exceeded

The following email could not be delivered as it exceeded the maximum size limit:

Subject: c++ books
Sender: safwanahaque@gmail.com

Please contact your IT Helpdesk or System Administrator if you require further assistance.

Fig. 3.Confusion matrix of phishing and ham mails



SNo	Code	Add Mismatch	Generic Salutation	Link	Link Length	IP	Dots	Domains	Click here	A	B	C	D	E	Urgent
1	p1	1	1	1	0	0	0	0	0	1	1	1	0	0	0
2	p2	0	0	1	1	0	0	0	0	0	0	0	0	0	0
3	p3	0	1	1	0	1	0	0	0	0	0	0	0	0	0
4	p4	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	p5	1	1	1	0	0	0	0	0	1	0	0	0	0	0
6	p6	1	1	1	1	0	0	1	1	0	0	1	0	0	0
7	p7	1	1	1	0	0	0	0	1	0	0	0	0	0	0
8	p8	1	1	1	1	0	0	0	0	0	0	1	0	0	0
9	p9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	p10	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 4.Confusion matrix of phishing and ham mails

The ‘extract features’ buttons were used to retrieve all the mail features used for this study, which were stored in a CSV file as shown in Figure 4.

The three components (database of files, GUI and machine learning classifier) described above were all designed and integrated to work together as a single unit to achieve the objectives of this study. The database consisted of the files that held the emails used for the study. These files were fed to the GUI, which provided the user interface and also to the classifier to carry out analysis. The JAVA segment extracted features, stored them in the database, which was then used to test and train the classifier. Feedbacks from users on their opinion on the type of mails they were viewing were collected through the GUI and saved in the database.

These results were then used in conjunction with the prediction of the classifier for comparison purposes. Upon detection of a phishing mail, the GUI alerts the user of the suspected features of the email. For easy understanding, brief characteristic outlines and diagrammatic relationships of the various components designed and implemented in this project are shown in Figure 5.

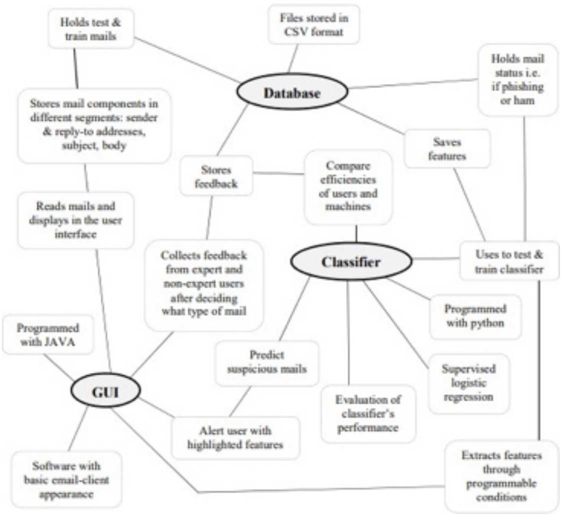

Holds test & train mails
Files stored in CSV format
Holds mail status i.e. if phishing or ham
Database
Stores mail components in different segments: sender & reply-to addresses, subject, body
Saves features
Stores feedback
Compare efficiencies of users and machines
Reads mails and displays in the user interface
Collects feedback from expert and non-expert users after deciding what type of mail
Classifier
Uses to test & train classifier
Programmed with python
Programmed with JAVA
Predict suspicious mails
Supervised logistic regression
GUI
Evaluation of classifier's performance
Alert user with highlighted features
Software with basic email-client appearance
Extracts features through programmable conditions

Fig. 5. Relationships between the different components of this project

III. Analysist and Result

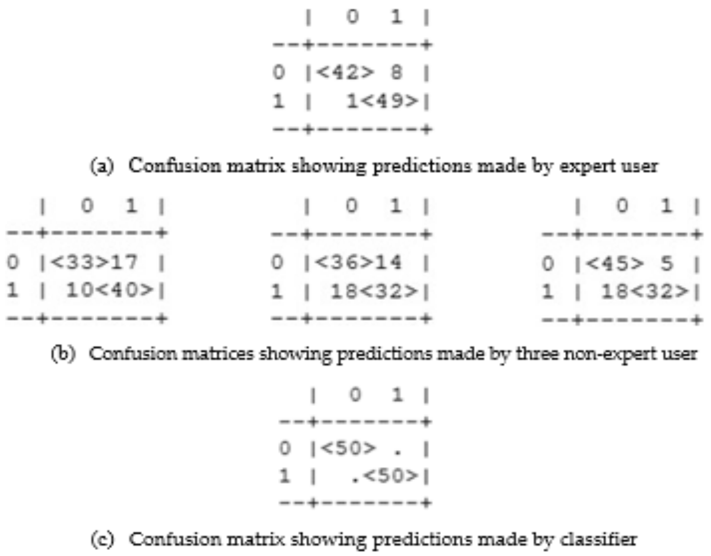
A. Confusion Matrix

The confusion matrix comprising true positives, false negatives, false positives and true negatives was determined for predictions gotten from the classifier, the expert user and the three non-expert users by comparing them to the actual truth values of the emails i.e. if ham or phishing. Figure 6(a), 6(b) and 6(c) show the confusion matrices of the predictions made by the expert user, non-expert users and the classifier respectively. In the figure, the rows represent the reference class while the columns represent the predicted class. 0 and 1 stand for ham and phish respectively.

From the confusion matrices;

- False positives alert the system of a phishing mail when the mail is not one resulting into false alarms. These errors may have drawbacks like excessive alerts for processing and sidelining legit mails.

- False negatives are the worst kind of errors as the predictions identify phishing mails as ham meaning the higher this value is, the higher the likelihood of falling victim to a phishing mail. Using the formula, FN/(TP+FN) to calculate the false negative rate (FNR), it can be seen that the expert user has an FNR of 2%, non-expert users have an FNR between 20%-36% and the classifier has a 0% FNR.



```
  |  0   1 |
--+--------+
0 |<42>  8 |
1 |   1<49>|
--+--------+
```

(a) Confusion matrix showing predictions made by expert user

```
  |  0   1 |        |  0   1 |        |  0   1 |
--+--------+      --+--------+      --+--------+
0 |<33>17  |      0 |<36>14  |      0 |<45>  5 |
1 |  10<40>|      1 |  18<32>|      1 |  18<32>|
--+--------+      --+--------+      --+--------+
```

(b) Confusion matrices showing predictions made by three non-expert user

```
  |  0   1 |
--+--------+
0 |<50>  . |
1 |    .<50>|
--+--------+
```

(c) Confusion matrix showing predictions made by classifier

Fig. 6. Confusion matrices of phishing and ham mails

B. Accuracy

The accuracy of predictions signifies how often the correct predictions are made i.e. true positives and true negatives. Thus, the higher the true negative and true positive values in a confusion matrix, the higher the accuracy of the overall prediction. From the confusion matrices in Figure 6, the computed prediction accuracies of the expert user is 91%, of the three non-expert users are 73%, 68% and 77% respectively and 100% for the classifier.

C. Precision, Recall, F-1 Score

Recall is used to determine how often true positive predictions are made out of all the actual positives in the sample while precision is the measure of the correctness of the predictions when true positives are predicted. F1-score is the mean of recall and precision values. Figure 7(a), 7(b) and 7(c) show the precision, recall and F-scores of the predictions made by the expert user, three non-expert users and the classifier.

D. ROC and AUC

ROC curve is a visual way of viewing the performance of a classifier. It is obtained by plotting the rate of making true positive predictions to the rate of making false positive predictions. The ROC curves of the expert user predictions, three non-expert user predictions and classifier predictions were all combined into a single graph for easy and clear comparisons as shown in Figure 8. The AUC is the area under curve and the value ranges between 0 to 1. The closer the value is to 1, the better the performance of the classifier considered. The AUC values are also shown in the legend of Figure 8.

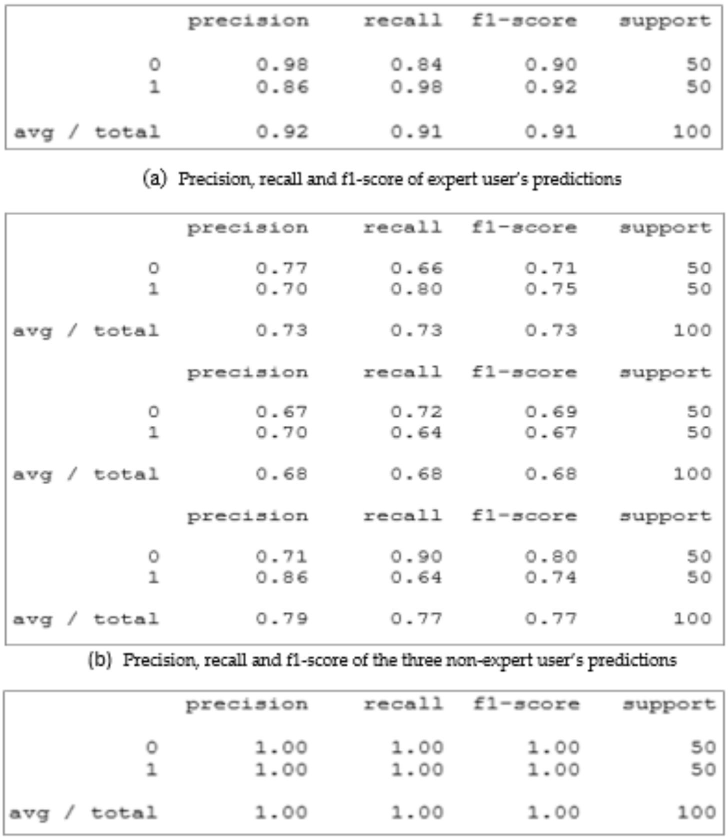
E. Reading and Writing to Files

In the design, there were numerous times when there was a need to access the database of files e.g. when feedbacks from the users were to be saved, when the mails from the files needed to be fed into the GUI and when phishing features of mails extracted by the programme needed to be stored in Features.CSV file.

F. After a GUI Phishing Mail Alert

After a possible phishing mail is predicted by the designed programme, an alert is issued to the user warning against the phishing mail and the possible phishing features that are contained in the mail. An example of when such a warning occurs is given in Figure 9. This warning occurs whenever a phishing mail is opened or viewed.

From the above results, it can be seen that the classifier has the best predictions compared to the expert and non-expert users. This shows that a machine learning classifier is very essential and useful in the application of detecting phishing mails as it minimises chances of falling victim to phishing attacks. It also demonstrates that an average user with little or no information technology (IT) knowledge or expertise will be at least 20% prone to phishing attacks. It can also be seen that the user with expert knowledge on phishing has a lesser chance of falling victim to phishing attacks compared to non-expert users. It can therefore be confidently said that the more educated and trained a person is on phishing mails and attacks, the better the person is equipped against these attacks and the lesser his/her chance of falling victim.



```
             precision    recall  f1-score   support

          0       0.98      0.84      0.90        50
          1       0.86      0.98      0.92        50

avg / total       0.92      0.91      0.91       100
```

(a) Precision, recall and f1-score of expert user's predictions

```
             precision    recall  f1-score   support

          0       0.77      0.66      0.71        50
          1       0.70      0.80      0.75        50

avg / total       0.73      0.73      0.73       100
             precision    recall  f1-score   support

          0       0.67      0.72      0.69        50
          1       0.70      0.64      0.67        50

avg / total       0.68      0.68      0.68       100
             precision    recall  f1-score   support

          0       0.71      0.90      0.80        50
          1       0.86      0.64      0.74        50

avg / total       0.79      0.77      0.77       100
```

(b) Precision, recall and f1-score of the three non-expert user's predictions

```
             precision    recall  f1-score   support

          0       1.00      1.00      1.00        50
          1       1.00      1.00      1.00        50

avg / total       1.00      1.00      1.00       100
```

(c) Precision, recall and f1-score of the classifier's predictions

Fig. 7. Precision, recall and f1-score of predictions

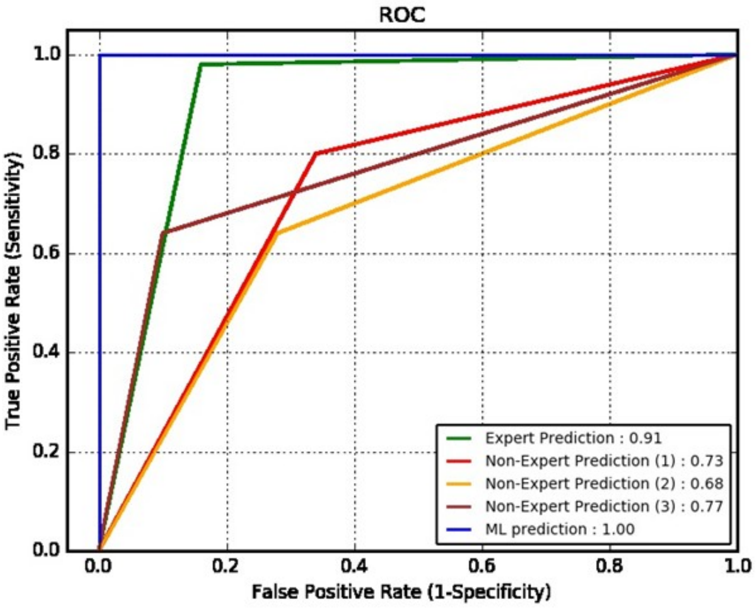
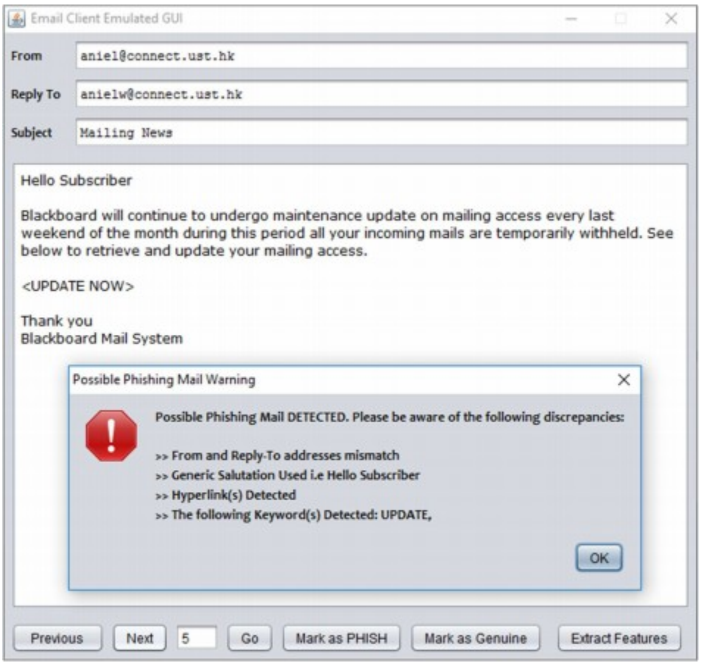


Fig. 8. ROC and AUC for predictions made by the expert user, three non-expert users and classifier



From: aniel@connect.ust.hk
Reply To: anielw@connect.ust.hk
Subject: Mailing News

Hello Subscriber

Blackboard will continue to undergo maintenance update on mailing access every last weekend of the month during this period all your incoming mails are temporarily withheld. See below to retrieve and update your mailing access.

<UPDATE NOW>

Thank you
Blackboard Mail System

Possible Phishing Mail DETECTED. Please be aware of the following discrepancies:

>> From and Reply-To addresses mismatch
>> Generic Salutation Used i.e Hello Subscriber
>> Hyperlink(s) Detected
>> The following Keyword(s) Detected: UPDATE,

5

Fig. 9. Warning issued when a possible phishing is suspected

IV. Conclusion

This study was carried out on one of the ever present and growing cybercrimes of this world known as phishing to address the current concerns of researchers and cyber security personnel, which are ways and techniques to phishing detection and understanding the human factor in phishing. This study thus, encompassed two main objectives, which were to detect a phishing mail with the prediction of a machine learning algorithm and also to study the ability of users to detect phishing mails successfully in relation to the predicting classifier. The responses or predictions of the classifier and users were categorised as true positives, false positives, true negatives and false negatives. These results were studied to compute the accuracies and efficiencies of the classifier and user predictions. For a phishing detection design, it is important that the false negative rate should be minimal as a higher rate indicates that more phishing attempts and attacks are going to go undetected and predicted as ham. The study showed that the machine learning algorithm had the lowest FNR of 0% meaning no phishing email was undetected. The expert user's prediction showed an impressive low FNR of 2% while the non-expert user predictions ranged between a high 20-36%. The study also showed a 100% overall prediction accuracy by the classifier and a very high 91% prediction accuracy by the expert user while the prediction accuracy of the non-expert users ranged between 68-77%. The null error rate of this classifier was 0.5 or 50%, which is the probability of being wrong if the majority of the class is always predicted (equal number of phishing and ham mails used for the experiment). The higher the value of a prediction from this value, the better the predicting mechanism. Comparing the accuracies to this null error rate, it can be seen that the non-expert users had a poor prediction capability. Hence, it could be deduced that a user's phishing prediction capability improves with the user's anti-phishing knowledge. The study also showed that an automated prediction for new and future mails had the highest accuracy and efficiency. All these results help to highlight and justify the need for automated phishing detection techniques and also the need to educate and train users on anti-phishing.

References

[1] Rekouche, K. Early Phishing, arXiv: 1106.4692, 2011.

[2] Kris, S. The Battle Against Identity Theft. Banker 2003, 153, 931.

[3] Eisenstein, E.M. Identity theft: An exploratory study with implications for marketers. Journal of Business Research 2008, 61, 1160–1172.

[4] Sullins, L.L. Phishing for a Solution: Domestic and International Approaches toDecreasing Online Identity Theft. Emory International Law Review 2006, 20, 397.

[5] Group, A.P. Phishing activity trends report, Anti Phishing WorkGroup, 1st Quarter, [online] Available at, 2016.

[6] Jung, J.S.; E.. An Empirical Study of Spam Traffic and the Use of DNSBlack Lists. (Accessed 2004, 18, 370–375.

[7] Felegyhazi, M.; Kreibich, C.P.; V.. On the Potential of Proactive. Domain Blacklisting., LEET 2010, 10, 6–6.

[8] Prakash, P.; Kumar, M.; Kompella, R.G.; M.. PhishNet: Predictive Blacklisting to. 2010, pp. 1–5.

[9] Dong, X.; Clark, J.J.; L, J. Defending the weakest link: phishing websites detection by analysing user behaviours. Telecommunication Systems 2010, 45, 215–226.

[10] Ramzan, Z. Phishing and Two-Factor Authentication Revisited, Symantec SecurityResponse, [online] Available, 2007.

[11] Nilsson, M.; Adams, A.H.; S..Building Security and Trust in OnlineBanking, In CHI '05 Extended Abstracts on; ACM: New York, NY, USA, 2005.1701–1704, [online] Available at: http://doi.acm.org/10.1145/1056808.1057001 (Accessed 17.

[12] Molloy, I.L.; N.. Attack on the GridCode One-time Password; ACM: New York, NY, USA, 2011; pp. 306–315. online] Available at:http://doi.acm.org/10.1145/1966913.1966953 (Accessed 17.

[13] Zviran, M.; Erlich.; Zippy. Identification and Authentication: Technology and Implementation Issues, Communications of the Association for Information Systems 2006, 17.

[14] Van Oorschot Mannan, M.; P.C.. Using a personal device tostrengthen password authentication from an untrusted computer; 2007; pp. 88–103. SpringerBerlin Heidelberg.

[15] Lu, H.P.; Lu, H.P.; Hsu, C.L.; Hsu, H.Y. An empirical study of the effect of perceived risk upon intention to use online applications. Information Management & Computer Security 2005, 13, 106–120.

[16] Plössl, K.; Federrath, H.; Nowey, T. Protection Mechanisms Against Phishing Attacks. Trust, Privacy, andSecurity in Digital Business, Lecture Notes in Computer Science, Springer Berlin Heidelberg

[17] DodgeJr., R.C.; Carver, C.F.; J, A. Phishing for user securityawareness. Computers & Security 2007, 26, 73–80.

[18] Kumaraguru, P.; Rhee, Y.; Sheng, S.; Hasan, S.; Acquisti, A.; Cranor, L.H. J.(2007) Getting Users to Pay Attention to Anti-phishing Education: Evaluation ofRetention and Transfer; ACM: New York, NY, USA, 2016; pp. 70–81.

[19] Robila, S.R.; W, J. Don'T Be a Phish: Steps in User Education; ACM: New York, NY, USA, 2006; pp. 237–241. Accessed 17 September 2016.

[20] Vishwanath, A.; Herath, T.; Chen, R.; Wang, J.R.; R, H. Why do people getphished? Testing individual differences in phishing vulnerability within anintegrated, information processing model. Decision Support Systems 2011, 51, 576–586.

[21] Del Castillo, M.D.; Iglesias, A.S.; I, J. Detecting Phishing E-mails by Heterogeneous Classification. Intelligent Data Engineering and Automated Learning IDEAL2007, Lecture Notes in Computer Science, Springer Berlin Heidelberg .

[22] Chandrasekaran, M.; Narayanan, K.U.; Phishing email detection based onstructural properties; 2006; pp. 1–7.

[23] Fette, I.; Sadeh, N.T.; A.. Learning to Detect Phishing Emails, InProceedings of the 16th International Conference on World Wide Web, WWW '07,New. (Accessed 2007, 18, 649–656.

[24] He, M.; Horng, S.J.; Fan, P.; Khan, M.K.; Run, R.S.; Lai, J.L.; andSutanto Chen, R.J.; A.. An efficient phishing webpage detector. Expert Systems withApplications 2011, 38, 12018–12027.

[25] Bergholz, A.; Beer, J.D.; Glahn, S.; Moens, M.F.; Paaß, G.S.; S.. Newfiltering approaches for phishing email. Journal of Computer Security 2010, 18, 7–35.

[26] Abu-Nimeh, S.; Nappa, D.; Wang, X.N.; S.. A Comparison of Machine LearningTechniques for Phishing Detection; Available: New York, NY, USA, ACM,pp. 60–69, [online, 2007].