Using Feature Selection and Classification Scheme for Automating Phishing Email Detection

Isredza Rahmi A HAMID[1], Jemal ABAWAJY[1], Tai-hoon KIM[2]

[1] School of Information Technology, Deakin University,
Waurn Ponds, VIC., 3217, Australia,
iraha@deakin.edu.au, Jemal@deakin.edu.au

[2] School of Computing and Information Science, University of Tasmania,
Centenary Building, room 350, Private Bag 87 Hobart TAS 7001 (Corresponding Author)
taihoonn@empas.com

Abstract: Email has become the critical communication medium for most organizations. Unfortunately, email-born attacks in computer networks are causing considerable economic losses worldwide. Exiting phishing email blocking appliances have little effect in weeding out the vast majority of phishing emails. At the same time, online criminals are becoming more dangerous and sophisticated. Phishing emails are more active than ever before and putting the average computer user and organizations at risk of significant data, brand and financial loss. In this paper, we propose a hybrid feature selection approach based on combination of content-based and behaviour-based. The approach could mine the attacker behaviour based on email header. On a publicly available test corpus, our hybrid features selection is able to achieve 94% accuracy rate.

Keywords: Internet Security, Behavior-based, Feature Selection, Phishing.

1. Introduction

Phishing emails have become common problem in recent years. According to Islam and Abawajy [22], "phishing attacks continue to pose serious risks for consumers and businesses as well as threatening global security and the economy". This calls for the development of effective countermeasures against email-born phishing attacks in order to safeguard critical infrastructures such as banking

Phishing is a type of semantic attack in which victims are sent emails that deceive them into providing sensitive information such as account numbers, passwords, or other personal to phisher. Normally, phishers send a large number of fake e-mails pretending to be from a legitimate and well-known business organization. Generally, the email content insists the victim to update personal information to avoid losing access rights to services provided by the organization.

Unfortunately, they lure user to a bogus web site implemented by the attacker. According to Anti-Phishing Working Group phishing trend report, the number of phishing attacks through email increased from about 170000 in 2005 to about 440000 in the 2009 [2]. Based on Gartner survey, approximately 109 million U.S adults have received phishing e-mail attacks with average loss per victim estimated to be $1,244.

Phishing email detection has drawn a lot of considerations from many researchers. Several good anti-techniques such as content-based [6], [11], [16] and behavior-based [7], [5], [13] have been developed to address the phishing problems. However, phishing attacks have continued to be a serious problem. This is because phishing has become more and more complicated and the phishes continually change their ways of perpetrating phishing attack to defeat the anti-phishing techniques. Moreover, most phishing emails are nearly identical to the normal email. Therefore existing anti-phishing techniques such as content-based approach are not able to curb phishing attacks. Furthermore, most of the existing emails filtering approaches are static where it can easily be defeated by modifying contents of emails and link strings.

In this paper, we present an approach to detect phishing email using hybrid features that combine content-based and behaviour-based approaches. The main objective of this paper is to identify behaviour-based features in phishing emails which cannot be disguised by an attacker. By analyzing attacker's pattern, it is observed that phishing email that has a tendency to come from more than one domain could indicate abnormal activity. Domain server that handles more than one type of domain email could show abnormal email as well. This information is done by analyzing email header which is usually neglected by others. We considered analyzing the message-ID tag and sender email in order to mine the attacker's behaviour. This study applies the proposed hybrid feature selection to 6923

datasets which come from Nazario [14] phishing email collection ranging from 2004 to 2007 and SpamAssassin [17] as ham emails. The result shows that the proposed hybrid feature selection approach is effective in identifying and classifying phishing email.

The remainder of this paper is organized as follows. Section 2 describes related research regarding phishing email detection approaches proposed in recent year. Section 3 examines the phishing email feature selection approach pertaining the data and feature set used in the experiment and hybrid feature selection algorithm as well. Section 4 gives the performance analysis result and the effectiveness of the proposed hybrid feature selection. Section 5 concludes the work and direction for future work is discussed.

2. Related Work

Several anti-phishing techniques have been proposed in recent years to detect and prevent the increasing number of phishing attacks. In general, phishing detection can be classified into server based techniques and client based techniques. Server based techniques typically are implemented by service providers such as ISP, e-commerce stores or other financial institutions. On the other hand, client-based techniques are implemented on users' end point through browser plug-ins or e-mail analysis.

El Ferchichi et al. [21] propose a wrapper approach to select features involving the Support Vector Machines (SVM) combined with a metaheuristic optimization algorithm: Tabu Search and Genetic Algorithms. The proposed process is based on the use of the rate of misclassification as an evaluating criterion. They apply the tabu algorithm to guide the search of the optimal set of features first and then a genetic algorithm is used to reach the same goal. Unlike our work which is focussed on phishing email detection, the work of El Ferchichi et al. [21] is applied on data from regulation of urban transport network systems.

Various feature selection approach have been recently introduced to assist phishing detection mechanism. Most of previous researches [6], [11], [16] were focusing on email content in order to classify the emails as either abnormal or normal. Previous attempt by [11] presents an approach based on natural structural characteristics in emails. The features included number of words in the email, the vocabulary, the structure of the subject line, and the presence of 18 keywords. They tested on 400 data which then divided into five sets with different type of feature selection. Their result shows the best when more features used to classify phishing email using Support Vector Machine classifier. However, the significance of the results is difficult to assess because of the small size of the email collection.

Fette et. al [6] on the other hand, considered 10 features which mostly examine URL and presence of JavaScript to flag emails as phishing. Nine features were extracted from the email and the last features obtained from WHOIS query. They follow similar approach as [11] but using larger datasets about 7000 normal emails and 860 phishing emails. They focused on URL properties which is not the best approach. This is because, attacker could use tools to obfuscate URL such as TinyUrl (http://tiny.cc/) and make it look valid. Their filter scores 97.6% F-measure and false positive rate of 0.13% and a false negative rate of 3.6% respectively.

Abu-Nimeh et al. [16] study the performance of different classifiers used in text mining such as logistic regression, classification and regression trees, Bayesian additive regression trees, Support Vector Machines, random forests, and neural networks. They test on a public collection of about 1700 phishing mails and 1700 legitimate mails from private mailboxes is used They focused on richness of word to classify phishing email based on 43 keywords. The features represent the frequency of "bag-of-words" that appear in phishing and legitimate emails. As phishing emails always look similar to normal email, this approach might not be reliable anymore.

Islam and Abawajy [22] propose a multi-tier phishing detection and filtering approach for phishing email filtering. They also propose a method for extracting the features of phishing email based on weighting of message content and message header and select the features according to priority ranking. The results of the experiments show that the proposed algorithm reduces the false positive problems substantially with lower complexity. Abawajy and Kelarev [23] propose a multi-tier ensemble construction of classifiers for phishing email detection and filtering.

Recently, behavior-based approach to determine phishing message has been proposed by [7], [5], [13]. Zhang et. al. [7] works on detecting abnormal mass mailing host in network layer by mining the traffic in session layer. Toolan et. al. [5] investigates 40 features that have been used in recent literature and proposed behavioral features such as number of word in send field, total number of characters in sender field, difference between sender's domain and reply-to domain and difference between sender's domains from the email's modal domain. Ahmed Syed at. al [13] however proposed behavioral blacklisting using 4 features which is log-based on live data. Ma et. al. [8] claimed they classify phishing email based on hybrid features. They used 7 features derived from 3 types of email features that are content feature, orthographic feature and derived feature which also can be considered as content-based approach as well.

In terms of detecting phishing using content, text-based classification does not seem to be the best approach. This is because phishing messages are nearly identical to the normal emails. Content-based filtering might be more effective technique if messages have a long lifetime and a large amount of duplication. However, attackers tend to use more sophisticated techniques from time to time that make them difficult to detect. They became more advanced to overcome this challenge by compiling phishing pages with non-HTML components, such as images, Flash objects, and Java applets. Yet, the updating rate of filters is often defeated by the changing rate of the attacks because phishing e-mails are continuously modifying senders and link strings. Therefore, this remains an open problem to be solved.

Although there are clear advantages to filtering phishing attacks at the email level, there are at present not many methods specifically designed to target phishing emails based on phishing behavior. There is a very little research on behavior-based approach. Our study differs from the previous work on feature selection in several ways. First we propose a hybrid feature selection by combining content-based and behavior-based features. We considered analyzing email header information particularly the sender email and email's message-ID tags in order to evaluate the attacker behaviors. We mine attacker behaviors by considering whether the sender sends emails from more than a single domain and if the domain name is used by more than one sender's domain. We then choose to use Bayes Net algorithm as our classifier because they are a powerful knowledge representation and reasoning mechanism. Second, we produce promising result using 7 features with 96% accuracy and 4% false positive and false negative rate respectively.

3. Feature Selection

Exiting email filtering approaches can be divided into origin-based filtering and content based filtering. Origin-based filtering focuses on the source of the e-mail and verifies whether this source is on a white verification list or black verification list. In contrast, content-based filters focus on the subject and body of the email. Phishing emails can be detected by filtering it based on text feature, linguistic feature or structural feature. The textual features and linguistic features identify phishing e-mails based on the word composition and grammatical construction. Instead, structural features focus on identifying the presence of obvious sign present in the e-mail body, which implicate it to be spoofed.

3.1 System model

An email message consists of three components, the message envelope, the message header, and the message body. The message header contains control information, including, sender's email address and one or more recipient addresses. There are other descriptive information is also added, such as a subject header field, message-id, a message submission date/time stamp and other information about the email.

When an email is sent, the message is routed from sender's server to the recipient's email server through MTA (Mail Transport Agent). MTA handles message transportation and acts as sorting area and mail carrier. This is where every email messages is stamped with email header information including message-id. This part of email header is not visible to most users but it is a useful indicator in determining phishing email. After that, MTAs communicate with one another using the SMTP protocol. The recipient's MTA then delivers the email to the MDA (Mail Delivery Agent) that acts as an incoming mail server. MDA is a mailbox where it stores the email as it waits for the user to

accept it. User then retrieve email using a software program called an MUA (Mail User Agent) such as Mozilla Thunderbird, Microsoft Outlook or Eudora Mail. Our feature selection approach could be deployed offline on the recipient's local machine as shown in Figure 1.

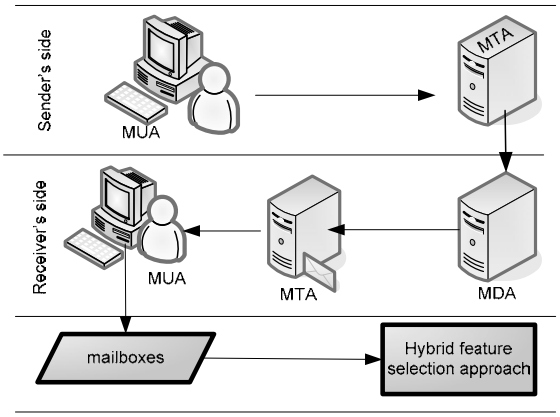

Figure 1. Hybrid Feature Selection Approach

Although different types of feature selection and classification algorithms for filtering phishing have been proposed in the literature, the scale and sophistication of phishing attacks have continued to increase steadily [22]. It is important to select the most relevant email features, which would contribute in increasing the performance of the detection algorithm by reducing dimensionality and processing time. This is because irrelevant and redundant features will impact the performances of classifiers and also slow the prediction process [20]. Thus appropriate email feature subset selection, which aims at choosing the most pertinent and representative features to increase accuracy rates and reliability of prediction models, is an essential step in the process of phishing email prediction.

3.2 Message-Id field validity

In this paper, a new behavior-based feature has been proposed which is based on information provided in the message-id field. The "Message-ID:" field is a unique message identifier that refers to a specific version of message. The uniqueness of the message identifier is guaranteed by the host that generates it. This message identifier is machine readable and not necessarily meaningful to humans. However, it could be used as an indicator to classify phishing email.

According to RFC2822 - Internet Message Format, Message-IDs have a specific format which is a subset of an email address and to be globally unique. Message ID consists of two parts, a local part and a domain, separated by an at-sign and enclosed in angle brackets: message-id: "<" local-part "@" domain ">". A common technique used by many message systems is to use a time and date stamp along with the local host's domain name, e.g., abc@example.com.

Each mail user agent (MUA) generates their own standard format of message-id field. Some of the attacker could forge the message-id field by deleting the message-id domain or change the domain name to make it look legitimate. Based on the attacker behavior, we analyze the message-id field whether the message-id value have been deleted or changed. We believe the phisher cannot modify the complete header, though he can forge certain fields. Therefore, email headers messages with blank message-id field and have uncommon domain name are considered as fake email. Common domain message-id name should have general top level domain such as .net, .com, .org and other registered domain.

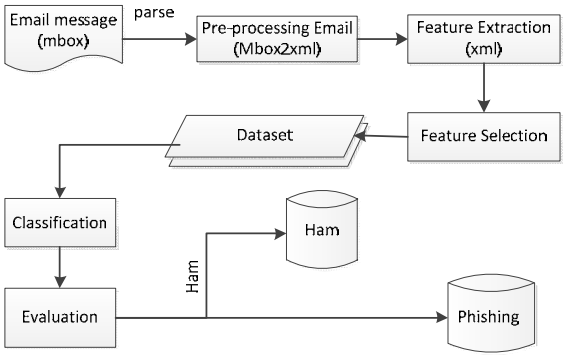

Figure 2. Hybrid Feature Selection System

3.3 Hybrid feature selection system

Figure 2 shows the basic system components and general processing steps which is extended from [15]. The processing phases includes: pre-processing of the email, feature extraction and selection, feature assessment, classification, and finally the evaluation of the classification result.

We used Bayes Net algorithm as our classifier as it is a powerful knowledge representation and reasoning mechanism. Moreover, it is the simplest and most widely used classification method because of its manipulating capabilities of tokens and associated probabilities according to the user's classification decisions and empirical performance.

We used open source software: Mbox2xml as a disassembly tool. A python module mbox2xml exported the information from mbox format to xml format. We modified some scheme in order to extract all features and store in the database. The next step in the process is to generate components of a feature vector by analyzing the database.

After that, we constructed 5 sets of datasets with various split percentage of ham and phishing emails. The training set was used to train the classifier and the test set to estimate the error rate of the train classifier. We use the same sets of data for training and testing the classifier as our main focused in this paper is to propose new behavior based feature selection approach.

3.4 Feature extraction & selection

It is well known that email consists of header and message body. Email header contains common identification such as from, to, date, subject and route information an email takes as it is transferred from one computer to another. It travels through a Mail Transfer Agent (MTA) where it is stamped with a date, time and recipient. This part of email header is not visible to most users but it is a useful indicator in determining phishing email. We find that message-ID tags found in email header is globally unique identification and can be used to mining the sender behavior.

The features that we identified in email header are: (1) Subject-based features: These features are related to the presence / absence of blacklist word in the email subject; (2) Sender-based features: These features are extracted from sender email address; (3) Behavior-based features: These features are extracted from the email header including information as sender email and email's message-ID.

The body-based feature includes the following: (1) URL-based: These features are extracted from email HTML; (2) Keyword-based: These features are related to the presence/absence of blacklist word in the email body; (3) Form-based: These features are related to the presence/absence of from in the email body; (4) Script-based: These features are related to the presence/absence of script in the email body.

3.5 Feature defines in email

Email messages have two basic parts that are the header and body parts. The header contains information about who the message was sent from, the recipients date and the route which contains optional fields such as received, reply-to, subject and message-ID. This is then followed by the body of the message. In our analysis, we considered the "message-ID" and the "from tag" in email header. We experimented with five features belong to email structure and additional two features which are extracted based on sender behavior. The features are listed as below:

1) Domain_sender: This binary feature represents the similarity of domain name extracted from email sender with domain message-ID. We think the email is normal if it is similar and set the value 0. If not, we set the value 1 to indicate the email is abnormal. This feature has been proposed by [5].
2) Subject_blacklist_words: This binary feature represents the appearance of blacklist words in the subject of an email which included in bags of words in [11]. If the email subject contains the blacklist word, the email is abnormal and set the value 1. This feature has been used in [8].
3) URL_IP: This numerical data shows number of links that are using IP address. This feature has been used in [1].
4) URL_dots: This numerical data represent number of links in email that contains dots more than 5. This feature has been used in [11] but they calculate maximum number of dots in every link.
5) URL_symbol : This numerical data represent the occurrence of links in emails that present symbol. This has been used in [18] but we incorporate other symbol such as "%" and "&"to detect obfuscation url.

The behavior features, proposed in this paper include:

6) Unique_sender (US): This binary data represent sender behavior whether the sender sends emails from more than a single domain. If it is more than 1, we think the sender is phisher and set value 1 or else the value is 0 to indicate that the sender is not phisher,
7) Unique_domain (UD): This binary data denotes if the domain names is used by more than one sender domain email. If it is more than 1, we think the email is abnormal or else the email is normal and set the value to 0, and

8) DMID_validity (DMID): the binary data denotes if the message-id field has been forged by the attacker. Note that this feature is an extension work from the one we discussed in [v]. Unlike the previous work that used 7 feature selection, in this paper the number of behavior-based feature selection has been increased.

3.6 Mining sender behaviour

The data mining for sender behaviour is analysed from email header. The dataset we selected from the email header has a structure as shown in Table 1.

Table 1. Datasets for sender behavior

Email Sender	Message-ID	Domain Message-ID	US	UD
service@paypal.com	q4-6$c--0--$-w@ qmb02q	qmb02q	1	0
service@paypal.com	YYBTIESSYZXKLGFQVFPTNKCG@hotmail.com	hotmail.com	1	0
mark@talios.com	2060000.1012684767@spawn.se7en.org	spawn.se7en.org	0	0
mark@talios.com	2060000.1012684767@spawn.se7en.org	spawn.se7en.org	0	0

Duration	Email messages
27 nov 2004 – 13 june 2005	414
14 june 2005 – 14 nov 2005	443
15 nov 2005 – 7 aug 2006	1423
7 aug 2006 – 7 aug 2007	2279
Total phishing datasets	4559

Table 2. Phishing datasets files summary

After all the features are defined, we extracted all 7 possible features from each email. The values of all features are in various types. Sender domain, subject blacklist word, unique sender and unique domain are in binary. All URL based features are in numerical however in vastly different ranges. For example, the URL dots could number of links under five. In order to treat all the original features as equally important, the value of each feature needs to be normalized before the classification process. Features with numerical values are normalized using the quotient of the actual value over the maximum value of that feature so that numerical values are limited to the range [0, 1].

3.7 Hybrid feature selection

In this section, we describe the proposed hybrid feature selection (HFS) algorithm. In the algorithm, we use domain_email_sender (DES), subject_blacklist_word (SBW), URL_dots (URLD), URL_symbol (URLS), URL_IP (URLIP), Unique_sender (US), Unique_domain (UD) and DMID_valid (DMID) feature values. Table III contains description commonly used notation in this algorithm. The HFS algorithm aims to determine feature matrix for predicting an email message is a phishing message or not. We then developed a methodology to extract seven features from each email [7].

First, the email messages is partitioned into four components containing ES, SE, MID and URL. The inputs to HFS algorithm are DES, SBW, URLD, URLS, URLIP, US, UD and DMID as shown in Figure 3. In step 1, reading count for each email is done. For step 2 to 5, each incoming emails will run functions to verify sender domain, identify email's subject blacklist word, URL feature matching and identify sender behaviour to extract features and finally construct the feature matrix. Unlike the previous algorithm, in this algorithm the number of behaviour-based feature selection has been increased where we proposed new feature: DMID_validity.

```
Algorithm HSF
 1: FOR (each incoming EMAIL) DO
 2:   FOR (i=1 to K) DO
 3:     Verify sender domain;
 4:     Identify blacklist word;
 5:     Perform URL feature matching;
 6:     Identify sender behavior;
 7:     Identify Message-id validity;
 8:     Constructing feature matrix;
 9:   ENDFOR
10: ENDFOR
END HSF
```

Figure 3. Hybrid Feature Selection Algorithm (HFS)

3.8 Identify Message-id Validity

Figure 2 shows the pseudo-code of the DMID algorithm. The input to the DMID is lists of domain message-id (DMID). In step 2 to 8, each incoming email will mine DMID value for all email messages to determine whether the email is phishing or normal email. If the

DMID's value has null value or contain uncommon generic top level domain name, the email is considered as forge email. DMID's value is set to 1 if it satisfies either condition.

```
Algorithm DMID
   INPUT: DMID
   SET US value to 0
   BEGIN
   1:    FOR (each incoming EMAIL) DO
   2:    FOR (i=1 to K) DO
   3:    IF (DMID[i] = = null + DMID[i] = =
         “”***.com","***.net","***.org","***.co","***.
         biz","***.edu","***.int","***.info") THEN
   4:       GIVE DMID value 1
   5: ELSE
   6: GIVE DMID value 0
   7: ENDIF
   8: ENDFOR
  END DMID
```

Figure 4. Algorithm for mining message-id validity

3.9 Constructing feature matrix

In this section, we construct the feature matrix of 8 features $F_i, i = 1,\cdots,8,$ for all phishing and normal datasets. Note that F_1, F_2, F_6, F_7 and F_8 are in binary while F_3, F_4 and F_5 are in numerical value ranging from 0 to 1. The R_i value for each features are summarized in Table 4.

4. Performance Analysis

4.1 Experimental Setup

This section presents our experimental setup. In our study, the classification was performed using WEKA (Waikato Environment for Knowledge Analysis). For our preliminary experiment, we used freely available pre-classified phishing datasets from [12]. We used 4 phishing dataset files as presented in Table 3. These phishing datasets have been used in phishing detection research including work by [3], [4], [5], [6], [9], [10], [12], and [16]. In order to provide non-phishing datasets, we used the SpamAssassin Project [17] from the easy ham directory. This collection provides 2364 hams emails.

We generated 5 sets of datasets randomly containing varying split percentage number of phishing and ham emails from the overall datasets. In order to treat the set equally, we fixed the number for each sets to 2000 data. The first set consists of 50:50 split percentage numbers of phishing email and ham email. The second set contains of 60:40 split percentages. The third set has 70:30 while the fourth set comprises of 80:20. Finally, the fifth set has the biggest percentage for ham email which is 90:10. Datasets which contained unreadable symbol, Chinese language and Nigerian online scam are neglected. The details on each datasets are summarized in Table 3.

Table 3. Summary of datasets for testing

Set	Percentage Ham : Phishing	Ham	Phishing	Total
1	50 : 50	1000	1000	2000
2	60 : 40	1200	800	2000
3	70 : 30	1400	600	2000
4	80: 20	1600	400	2000
5	90 : 10	1800	200	2000

4.2 Performance Metric

In order to measure the effectiveness of the classification, we refer to the four possible outcomes as:

1. *True positive (TP):* a classifier correctly identifies an instance as being positive.
2. *False positive (FP):* a classifier incorrectly identified an instance as being negative, in fact an instance is instances hypothetical to be positive.
3. *True negative (TN):* a classifier correctly identifies an instance as being negative;
4. *False negative (FN):* a classifier incorrectly identifies an instance as being positive, in fact an instances hypothetical to be negative.

Table 4. Classification result of 5 data sets

Set	TP	FN	FP	TN
1	0.92	0.08	0.08	0.92
2	0.94	0.06	0.07	0.94
3	0.95	0.06	0.07	0.93
4	0.95	0.06	0.15	0.85
5	0.97	0.03	0.26	0.74

(a)

Set	ACC	pre	err	Recall
1	92%	0.92	0.08	0.923
2	94%	0.94	0.06	0.939
3	94%	0.93	0.06	0.945
4	90%	0.86	0.10	0.945
5	86%	0.79	0.14	0.969

(b)

To measure the effectiveness of our approach, we use four metrics that are also used in previous work [5], [6], [8] and [11]:

5. *Precision (P)* - this is the fraction of correctness;
6. *Recall (R)* - this measures the portion of the completeness of correct categories that were assigned;
7. *Accuracy (A)* - this measures the percentage of all decisions that were correct; and
8. *Error (E)* - this relates to the number of misclassifications of instances.

4.3 Results and Discussions

This section presents the classification outcome of the Bayes Net algorithms on the extracted features. We decided to test the feature selection approach using Simulated Annealing search algorithm with 10 folds cross validations.

4.3.1 Feature Selection

Table 4 (a) and (b) presents the experimental results according to selected classifier for five sets of data. Our result shows that, the hybrid based feature selection by combining content-based and behaviour-based feature selection shows quite promising result. This is evidence that features based on sender and domain behaviour could be considered to determine phishing email.

We tested on 5 sets of data with various split percentages of phishing and ham messages. Data for set 2 and set 3 achieved the highest accuracy. In contrast, data set 5 showed the lowest accuracy among other datasets.

4.3.2 Comparative Analysis

In Table 5, we compare our result with existing works that used the same dataset from [12] and achieving at least 80% accuracy. Fette et. al. [6] proposed 10 features mostly based on URL and script presence achieved 96% accuracy. They used Random Forest as a classifier.

Abu Nimeh [15] examined 43 keywords generated using TF-IDF (Term Frequency-Inverse Document Frequency) as an indicator to determine the best machine learning technique for phishing email detection. They compare the accuracy between several machine learning methods including Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NNet) for predicting phishing emails. They found that Neural Net algorithm performs the best among others with 94.5% accuracy.

Toolan et. al [5] used content-based and behaviour-based approach to classify phishing email similar to the one describe in the current paper. They used 22 features to test on 3 datasets comprising 6097 samples. They achieved approximately 97% accuracy.

Finally, we include our work who aimed to proposed hybrid feature selection using 8 features. We successfully achieved 94% accuracy covering 6923 samples. Even though the accuracy is quite low, we manage to test it only by using more robust features and least feature selection compared to others.

Table 5. Comparison of the approaches

	Feature Approach	**Sample**	**Accuracy**
Fette et. al [6]	URL-based and script-based.	7810	96%
Abu-Nimeh et.al. [15]	Keyword-based.	2889	NN (94.5%) RF (94.4%) SVM (94%) LR (93.8%) BART (93.2%) CART (91.6%)
Toolan et. al. [5]	Behavioral-based and content-based.	6097	Dataset 1 (97%) Dataset 2 (84%) Dataset 3 (79%)
Ours	Hybrid feature.	6923	Set 1 (91%) Set 2 (93%) Set 3 (92%)

4.3.3. Other finding

Experiments were conducted with four different type of classification algorithm to identify which machine learning method performs the best. We have implemented using Bayes Net, support vector machine (SVM), AdaBoost and Random Tree.

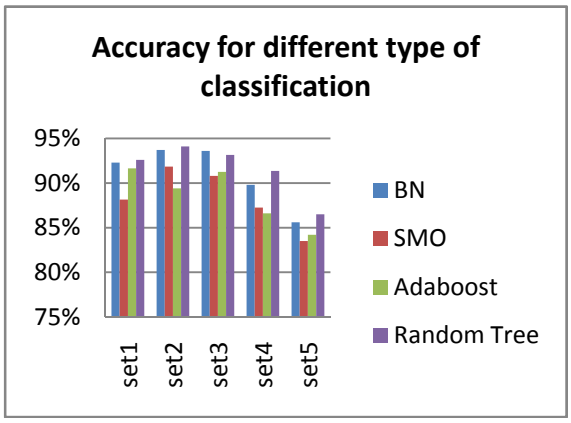

Figure 5. Accuracies for different type of classification

Figure 5 shows that the result comes that Bayes Net generated the highest accuracy which builds a good classifier. Comparing to other classification algorithm, the highest accuracies of other classification algorithms are AdaBoost (-0.02%), Support Vector Machine (-0.02%), and Random Tree (0.00%). This result recommends that Bayes Net and Random Tree achieved the highest accuracy and work well in discrete and small vector space data.

5. Discussion

In this paper, we propose behavior-based features to detect phishing emails by observing sender behavior. We extract all features using Mbox2xml as a disassembly tool. We then mine the sender behavior to identify whether the email came from legitimate sender or not. We take into account behavior of sender who tends to send email from more than a single domain and a domain that handle different kind of email sender domain. Other than that, the attacker also used to forge the message-id field information to cover their tracks.

By combining these datasets, we used Bayes Net algorithm to classify the datasets into phishing or ham emails. This hybrid feature selection approach produce promising result using 8 features with 94% accuracy. The feature selection we used in this paper does not work on graphical form as some attacker bypass the content based approach using image. The result motivates future works to explore attackers' behaviour and profile their modus operandi. As future works, we would like to investigate further on message-id field to understand the attacker strategies to cover their tracks.

Acknowledgements

This paper will not be completed without the help of Maliha Omer.

REFERENCES

1. BERGHOLZ, A., G. PAAB, F. REICHARTZ, S. STROBEL, J. H. CHUNG, **Improved Phishing Detection using Model-based Features**, In Proc. of the International Conference on E-mail and Anti-Spam, 2008.
2. The Anti-Phishing work Group. Available: http://www.apwg.org/
3. LIU, C., S. STAMM, **Fighting Unicode-Obfuscated Spam**, Proceedings of E-Crime Research, ACM, New York, USA, 2007, pp. 45-59.
4. TOOLAN, F., J. CARTHY, **Phishing Detection using Classifier Ensemble**, In eCrime Researchers Summit, 2009, pp. 1-9.
5. TOOLAN, F., J. CARTHY, **Feature Selection for Spam and Phishing Detection**, In eCrime Researchers Summit (eCrime), 2010, pp. 1-12.
6. FETTE, I., N. SADEH, A. TOMASIC, **Learning to Detect Phishing Emails**, Proceedings of the 16th International Conference on World Wide Web (WWW '07), ACM, New York, USA, 2006, pp. 649-656.
7. ZHANG, J., Z. DU, W. LIU, **A Behaviour-based Detection Approach to Mass-Mailing Host**, In Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, vol. 4, 2007, pp. 2140-2144.
8. MA, L., B. OFOGHANI, P. WATTERS, S. BROWN, **Detecting Phishing Emails using Hybrid Features**, Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, 2009, pp. 493-497.

9. ZHOU, L., Y. SHI, D. ZHANG, **A Statistical Language Modelling Approach to Online Deception Detection**, IEEE Transactions on Knowledge and Data Engineering, vol. 20, No. 8, 2007, pp. 1077-1081.

10. BAZARGANIGILANI, M., **Phishing E-Mail Detection using Ontology Concept and Naïve Bayes Algorithm**, International Journal of Research and Reviews in Computer Science (IJRRCS), vol. 2, no. 2, 2011, pp. 249-252.

11. CHANDRASEKARAN, M., K. NARAYANAN, S. UPADYAYA, **Phishing Email Detection Based on Structural Properties**, Proceeding of the Cyber Security Conference, 2006.

12. CHANDRASEKARAN, M., V. SHANKARANARAYANAN, S. UPADHYAYA, **CUSP: Customizable and Usable Spam Filters for Detecting Phishing Emails**, Proceeding 3r Annual Symposium on Information Assurance (ASIA '08), Albany, NY, 2008, pp. 10-17.

13. SYED, N. A., N. FEAMSTER, A. GRAY, **Learning To Predict Bad Behaviour**, NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security, 2008.

14. NAZARIO, J., **Phishing Corpus**, Available: http://www.monkey.org/jose/wiki/doku.php?id=phishingcorpus.

15. BASNET, R. B., A. H. SUNG, **Classifying Phishing Emails using Confidence-Weighted Linear Classifiers**, International Conference on Information Security and Artificial Intelligence (ISAI 2010), 2010, pp. 108-112.

16. ABU-NIMEH, S., D. NAPPA, X. WANG, S. NAIR, **Comparison of Machine Learning Techniques for Phishing Detection**, Proceeding of APWG eCrime Researchers Summit, Pittsburgh, ACM, New York, USA, 2007, pp. 60-69.

17. **Spamassassin public corpus**, Available: http://spamassassin.apache.org/publiccorpus.

18. GANSTERER, W. N. D. POLZ, E-Mail **Classification for Phishing Defence**, in LNCS Advances, Volume 5478, 2009, pp 449-460.

19. A HAMID, I. R., J. H. ABAWAJY, **Hybrid Feature Selection for Phishing Email Detection**, The 11th International Conference on Algorithms and Architectures for Parallel Processing, Springer, Berlin, Germany, 2011, pp. 266-275.

20. PENG, Y., G. KOU, D. ERGU, W. WU, Y. SHI, **An Integrated Feature Selection and Classification Scheme**, Studies in Informatics and Control, ISSN 1220-1766, vol. 21 (3), 2012, pp. 241-248.

21. FERCHICHI, S. E., K. LAABIDI, S. ZIDI, **Genetic Algorithm and Tabu Search for Feature Selection**, Studies in Informatics and Control, ISSN 1220-1766, vol. 18 (2), 2009, pp. 181-187.

22. ISLAM, R., J. H. ABAWAJY, **A Multi-tier Phishing Detection and Filtering Approach**, Journal of Network and Computer Applications, vol. 36 (1), 2013, pp. 324-336.