Detecting phishing e-mails using Text and Data mining

Mayank Pandey , Vadlamani Ravi*
Institute for Development & Research in Banking Technology, Masab Tank, Hyderabad, India
*Corresponding Author
(mayank08p@gmail..com , rav_padma@yahoo.com)

Abstract - **This paper presents text and data mining in tandem to detect the phishing email. The study employs Multilayer Perceptron (MLP), Decision Trees (DT), Support Vector Machine (SVM), Group Method of Data Handling (GMDH), Probabilistic Neural Net (PNN), Genetic Programming (GP) and Logistic Regression (LR) for classification. A dataset of 2500 phishing and non phishing emails is analyzed after extracting 23 keywords from the email bodies using text mining from the original dataset. Further, we selected 12 most important features using t-statistic based feature selection. Here, we did not find statistically significant difference in sensitivity as indicated by t-test at 1% level of significance, both with and without feature selection across all techniques except PNN. Since, the GP and DT are not statistically significantly different either with or without feature selection at 1% level of significance, DT should be preferred because it yields 'if-then' rules, thereby increasing the comprehensibility of the system.**

Keywords - **Multilayer Perceptron, Decision Tree, Logistic regression, Support Vector Machine, Group Method Of Data Handling, Phishing webpage, Probabilistic Neural Network, Genetic Programming, Text mining, Classification.**

I. INTRODUCTION

Phishing is the major security threat in the internet society. According to one survey by anti phishing working group [1], there were 83,083 unique phishing attacks worldwide, in 200 top-level domains. Phishing is the act of sending an email to a user falsely claiming to be an established legitimate enterprise in an attempt to scam the user into surrendering private information that will be used for identity theft [2]. The phishing emails redirect the user to a website, which is a look-alike of the legitimate site, where they are asked to fill their credentials such as password, bank account number, social security number, credit card details that the legitimate site already has. Phishers use various methods to attract the web users, such as they send greeting to users and send menace messages that give indication to update account [8],[17]. Phishing attacks are bigger threat in banking and financial institution; this is the area where they target to attack. To rescue users from phishing emails, we need to develop a mechanism that should be capable of identifying phishing e-mails even if it is new [10]. The motivation behind this research is to make effective predictions to detect phishing emails using machine learning techniques.

Various researches have worked on phishing e-mails detection. Many of them considered URL as the important component to extract features, but in this work, our focus is on the content of the e-mail.

The rest of the paper is structured as follows: Literature review is presented in Section II. Proposed methodology is presented in section III, followed by results and discussion in section IV and finally conclusion in section V.

II. LITERATURE REVIEW

A lot of work has been reported in the field of detecting phishing e-mails. Most of them focused on the URL part of the phishing e-mails. Our study is based purely on the content part of phishing e-mails and not on the URL. This section contains the study of related works.

For phishing detection Saeed et al. [5] proposed model using machine learning techniques. They used LR, Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), SVM, Random Forests (RF) and NN. The dataset used by them contains 2889 emails of which 59.5% are legitimate used from [6] and the rest are phishing e-mails taken from [7]. They concluded that RF yielded highest accuracy of 92.28%.

Then, Lakshmi et al.,[8] used 17 features extracted from source code of phishing URLs orweb sites, taken from Phishtank [9], including 100 phishing and 100 legitimate sites. the employed MLP, Decision Tree (J48) and NB. They obtained highest accuracy of 98.5% using J48. This study is an extension of Adi et al. [10] who obtained the accuracy of 97.33%. A similar approach is followed by [16] where the authors selected lexical features of URL.They obtained 98% accuracy from their research.

Basnet et al. [11] selected 16 features from various parts of phishing e-mail URL and some keywords of phishing webpage. The dataset used in the study is from Phishing Corpus [7] for Phishing e-mails and Spam Assassin [12] for legitimate e-mails. The dataset has 4000 e-mails, with 973 phishing e-mails. They used Biased

978-1-4673-1344-5/12/$31.00 ©2012 IEEE

Support Vector Machine (BSVMs), SVMs, NN and K-Means. They obtained highest accuracy of 97.99% with BSVM and NN.

Maher et al. [18] used fuzzy inference for evaluating e-mail phishing rate. They got worst website phishing rate 86.2% representing very phihsy website and best 13.8% representing legitimate website.

Ammar et al. [19] used fuzzy neural network, for phishing e-mail detection. They used 16 features extracted from HTML part of e-mails, IP, and different characteristics of URL. The dataset used is collected from [7] for phishing e-mails and [12] for legitimate e-mails. They used Phishing Evolving Neural Fuzzy Framework technique and compared its performance with that of MLP and MLR. They obtained best error rates of 0.13 and 0.12.

Sadia and Rachel [20] proposed an automated web phishing detection approach, called PhishZoo, using profiling and Fuzzy matching. They explained problems of blacklist and whitelist approach [21,22,23]. They obtained sensitivity of 97.14% using PhishZoo, which is higher compared to Netcraft [24] and Firefox version 3 [25].

Fergus, Joe [17] obtained very good results with Classifier ensembles. Using individual techniques, they obtained 97.15% accuracy by C5.0. With classifier ensembles they obtained 93.68% accuracy but they achieved better recall using classifier ensembles than individual techniques.

Maher et al. [15] proposed fuzzy data mining for detecting e-banking phishing websites. They got 83.7% phishing website rate representing very phishy website and 16.4% representing legitimate website using all three layers with input value. Our motivation behind this research was to generate an effective classification model for phishing e-mail detection using less number of features. Most of the previous researchers has developed phishing detection model based on URL, Source code and IP related features but in our study we select only structure part of phishing emails. Here, we performed feature selection and compare the results of both methods. We used the same dataset which has been used by previous researchers [7], [12].

The present study is different from the previous ones as follows: (i) We constructed the dataset by selecting features from only the body part of the emails, which is very efficient and easy way in comparison to selecting features from URL & source code of the websites. (ii) Further, we performed feature selection using t-statistic.

III. PROPOSED METHODOLOGY

A. Data Collection and Preparation

To detect the phishing e-mail, we analyzed the dataset with 2500 e-mails in which 1260 are phishing e-mails and remaining 1240 are legitimate. The phishing dataset is constructed by processing sample phishing emails collected from Phishing Corpus [6]. The legitimate part of dataset is constructed by processing legitimate emails collected from Spam Assassin [12]. We selected 23 features which are related to suspected terminology based on most frequently occurring keywords in the web pages. Such dataset is not available on internet so we have to prepare it. We collected unstructured textual data from phishing and legitimate e-mails. Then, we converted it into appropriate structured data through data preparation and finally prepare dataset using text mining. Text mining is performed using the tool Rapid Miner [13]. Rapid miner provides a list of all unique words present in the data and along with their frequency. We selected those keywords which have high frequency and relevant to the phishing problem and labeled them as features. The details are presented in the sub-section C below. Fig. 1 depicts the overall methodology for detecting phishing emails.

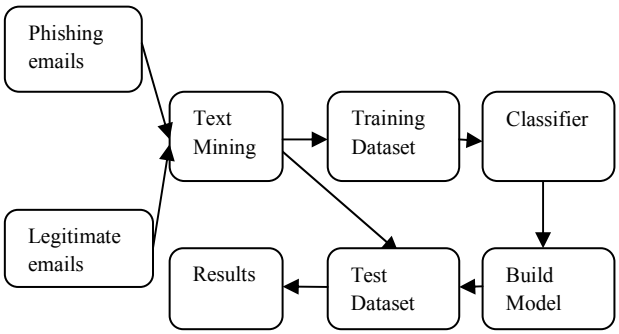

Fig. 1. Methodology for detecting phishing emails

B. Overview of the Techniques

1). Group Method of Data Handling (GMDH)

This is a family of inductive algorithms for mathematical modeling of multi-parametric datasets that features fully-automatic structural and parametric optimization of models. GMDH can find relations in data to select optimal structure of model or network or to increase the accuracy of existing algorithms. This self-organizing approach is different from commonly used deductive modeling. It is inductive as the best solution is found by sorting-out of possible variants and the algorithm itself finds the structure of the model and the laws of the system (http://en.wikipedia.org/wiki/GMDH).

GMDH algorithms inductively sort out gradually complicated polynomial models and select the best solution by means of the *external criterion*. A GMDH

model with multiple inputs and one output is a subset of components of the *base function*, as in (1).

$$Y(x_1, \dots, x_n) = a_0 + \sum_{i=1}^{m} a_i f_i. \quad (1)$$

Where, f_i are elementary functions dependent on different sets of inputs, a_i are the coefficients and m is the number of the base function components. GMDH algorithm considers various component subsets of the base function called *partial models* and the coefficients of these models are estimated by the least squares method. The number of partial model components is gradually increased to find a model structure with optimal complexity indicated by the minimum value of an *external criterion*. This process is called self-organization of models [27].

GMDH is also known as polynomial neural networks and statistical learning networks due to implementation of the corresponding algorithms in several commercial software products (http://en.wikipedia.org/wiki/GMDH).

2). Genetic Programming

Genetic programming (GP) is a biologically inspired evolutionary algorithm to find computer programs that perform a given task. It is like genetic algorithms (GA) but here each individual is a computer program. It optimizes a population of computer programs according to a fitness landscape based on a program's ability to perform a given computational task.

Being computationally intensive in the 1990s GP was mainly used to solve relatively simple problems. But thanks to improvements in GP algorithms and to the exponential growth in CPU power, GP has become more prevalent and has produced many novel and outstanding results in areas such as quantum computing, electronic design, game playing, sorting, searching etc (http://en.wikipedia.org/wiki/Genetic_programming). GP evolves computer programs, represented in memory as tree structures which can be easily evaluated recursively (http://en.wikipedia.org/wiki/Genetic_programming).

Every tree node has an operator function and every terminal node has an operand, making mathematical expressions easy to evolve and evaluate. The main operators used in GP are crossover and mutation. Crossover is applied on an individual by switching one of its nodes with another node from another individual in the population. With tree-based representation, replacing a node means replacing the whole branch which gives greater effectiveness to the crossover operator. The *children* expressions resulting from crossover are very much different from their initial parents. Mutation affects an individual in the population. It can replace a whole node in the selected individual, or it can replace just the node's information [27].

C. Text Mining

Text mining is a new and exciting research area that tries to solve the information overload problem by using techniques from the data mining, machine learning, natural language processing, information retrieval and knowledge management. Text mining involves the preprocessing of document collections, the storage of the intermediate representation, the techniques to analyze these intermediate representation and visualization of results [3, 26]. Text mining algorithm works on extracting the hidden pattern on textual data. It is used to convert the textual data into numeric dataset. Text-mining algorithm works on feature based characterization of the documents. Characters, words, terms and concepts are the potential features used to represent the documents. It gained popularity as a research area due to its ability to mine unstructured/digital content available on the internet [4].

To prepare the dataset from textual data we used term-document matrix. We selected those keywords which comes frequently and related to phishing terminology. In the process of selecting keywords we grouped most occurring similar meaning words into one keyword such as ‘account holder’, ‘customer’, ‘member’ can be grouped as one keyword ‘member’. Various tasks are involved in text mining such as text tokenization, stemming, filtering and segmentation. We extracted keywords from document and used them as features in input dataset to classify the phishing and legitimate emails.

D. Document -Term Matrix

In this research we used Document -Term Matrix approach for constructing our experimental dataset features from collection of phishing and legitimate e-mails. Table I presents an example of the Document – Term Matrix. Let D1 and D2 be two documents and Phishing, PayPal, Attacks and Threat be the terms.

D1: Phishing is major security threat
D2: PayPal is mostly used for phishing attacks

TABLE I
DOCUMENT TERM MATRIX

	Phishing	Paypal	Attacks	Threat
D1	1	0	0	1
D2	1	1	1	0

It is a matrix that contains the frequency of every term in a collection of documents. This approach falls under text mining task. In the matrix, rows represent the

documents and columns represent the selected words or terms of document. In this study, we selected the most frequently occurring terms in the emails as features which are relevant to the phishing detection.

E. Features Used

In this research, 23 features are used for identifying the phishing emails. Structural keywords are used as features and we extract them based on their frequency of occurrence and relevance to the phishing term. From the collected emails, we extracted those keywords such as account, transaction, credit etc., which are mostly used in the websites where sharing of credential information and transactions occur. Table II presents the list of all keywords used as features in this study**.** We trust that the extracted features would be able to identify a phishing email.

F. Feature Selection

One of the main issues in classification problems is the identification of key features that improve the prediction accuracy. It is possible to identify and remove the redundant features, which do not contribute to the improvement of the classification accuracy. In this research, we used t-statistic method to perform feature selection. Using this technique we computed t- statistic values for each feature of the dataset and ranked the features in the descending order of the t-statistic value. Higher the t-statistic value, more discriminating is the feature said to be. Accordingly, we selected top 50% features viz. top 12 features for classification purpose. Thus, the new input dataset with the selected features is fed to the classification models. In Table II highlighted keywords are the most important ones, selected by the feature selection process.

The t-statistic is one of the effective techniques used for feature selection [14]. Equation (2) is the formula for computing t-value. This technique is used for only binary classification problem.

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (2)$$

where μ_1 & μ_2 represent the means for two different classes of dataset, σ_1 & σ_2 represent standard deviations for each class and n_1 & n_2 represent number of samples in each class.

In this research, we used 10-fold cross validation method to test the classification models. We used data mining techniques viz., LR, DT, MLP, SVM, GMDH, PNN, GP as classifiers and performed classification task for identifying the phishing emails.

IV. RESULTS & DISCUSSION

We conducted two experiments with the same dataset, first without feature selection and then with feature selection. Sensitivity and accuracy is used to evaluate the performance of the classifiers.

We used DT, LR, MLP, SVM, PNN, GMDH and GP as classifiers. DT, LR, SVM, MLP, PNN are implemented using the tool Knime (http://www.knime.org/), GMDH is implemented using Neuroshell (http://www.neuroshell.com/) and GP is implemented using Discipulus tool (http://www.rmltech.com/). The average results of 10 folds using above classification techniques are presented in Table III. We compared the results of with and without feature selection of dataset.

TABLE II
KEYWORDS EXTRACTED AND THEIR t-STATISTIC VALUE

Keyword	t-statistic value	Keyword	t-statistic value
Account	42.11758	Response	4.504742
Member	19.31017	Offer	3.3573
Access	17.05149	Transaction	2.809165
Email	16.6749	Agreement	1.863823
Address	15.3276	Registration	1.791679
Update	15.03956	Person	1.522086
Price	12.55556	System	1.276965
Market	10.23963	Process	1.076627
Online	9.770984	Service	0.895862
Information	9.474589	Request	0.616316
Work	8.331161	Message	0.32805
Credit	5.883174		

Using 23 features the GP yields the best results in terms of accuracy and sensitivity. We obtained 98.12, 97.29 as accuracy and Sensitivity respectively using GP followed by MLP, DT, GMDH, SVM, LR and PNN in terms of accuracy. In comparison with Saeed et al.[5], Ram et al. [11], Fergus and Joe [17] and Madhusudhanan et al. [28], we obtained better accuracy for detecting the phishing emails. Table III presents the accuracy, sensitivity and specificity of various classifiers with and without feature selection.

Using t-statistic based feature selection we selected 12 features out of the original 23 and fed them to the classifiers. The results obtained with and without feature selection are very much similar. Here also, the GP yielded the best accuracy and sensitivity of 97.6, 96.82 respectively followed by MLP, DT, SVM, GMDH, LR and PNN in terms of accuracy.

TABLE III
AVERAGE 10-FOLD RESULTS OF ALL CLASSIFIERS AND THEIR RESPECTIVE t-STATISTIC VALUE

Classifier	Without Feature Selection			With Feature Selection			t-statistic value (Sensitivity)
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	
GP	**98.12**	**97.29**	98.95	**97.6**	**96.82**	**98.38**	0.684
MLP	97.2	97.12	97.28	96.72	96.49	96.94	0.907
DT	95.8	96.09	95.5	95.4	95.8	95.01	0.204
GMDH	95.12	94.37	94.36	94.36	93.8	94.92	0.379
SVM	94.24	92.86	95.67	94.36	93.81	94.9	0.727
LR	93.48	91.6	95.41	94	92.62	95.42	0.67
PNN	91.84	94.66	88.94	92.24	96.8	87.58	3.072

By comparing the results of with and without feature selection it turned out that they yielded approximately the same results. Hence, we conducted t-test to see if the difference between sensitivities in both cases is statistically significant. The t-statistic values computed for DT, LR, SVM, MLP, GMDH, PNN and GP are 0.204, 0.67, 0.727, 0.907, 0.379, 3.072, and 0.684 respectively. The t-value is less than 2.83 (which is the t-table value at 10+10-2=18 degrees of freedom) for all classifiers except PNN, where feature selection indeed improved the sensitivity of PNN. It means in all classifiers except PNN, using only 12 features, we are able to get statistically the same sensitivity, which is a significant result of the study.

TABLE IV
Model Comparison Based On t-Test

Classifiers	t-statistic value (Sensitivity)	
	Without FS	With FS
GP vs. DT	1.25	1.12
GP vs. LR	5.30	3.27
GP vs. SVM	4.31	2.83
GP vs. MLP	0.25	0.47
GP vs. GMDH	2.41	2.74
GP vs PNN	3.48	0.03

Further, since GP yielded numerically best results, we compared it with all other classifiers in both cases of with and without feature selection. Thus, we performed t-test again on GP versus other classifiers. Table IV presents the model comparison in both the cases. Here, in the case of without feature selection, we found that DT, MLP & GMDH are not statistically significantly different in comparison with GP. However, compared to LR, SVM & PNN, GP is statistically significantly better. Then, with feature selection, we found that except LR all the classifiers are not statistically significantly different compared to GP. Therefore, an important insight found here is that feature selection did not worsen the results of majority of the classifiers in detecting the phishing emails. Since, the GP and DT are not statistically significantly different either with or without feature selection, we infer that DT should be preferred because, it is not a black box in that it yields 'if-then' rules, which can act like an early warning expert system.

V. CONCLUSION

Phishing is one of the most dangerous cyber attacks and the correct detection of phishing attacks has become a major challenge for users. In our research, we developed text and data mining based techniques to detect the phishing emails. A dataset of 2500 phishing and non phishing emails is analyzed after extracting 23 keywords from the email bodies using text mining from the original dataset. Further, we selected 12 most important features using t-statistic based feature selection. Here, we conducted experiments with and without feature selection. The primary objective behind this research was to achieve higher phishing prediction accuracy with less number of features. We found that feature selection did not worsen the results of majority of the classifiers in detecting the phishing emails. Since, the GP and DT are not statistically significantly different either with or without feature selection, we infer that DT should be preferred because, it is not a black box in that it yields 'if-then' rules, which can act like an early warning expert system.

ACKNOWLEDGEMENT

We are thankful to Mr. Frank Francone for permitting us to use Discipulus Tool (Demo version) for conducting experiments involving genetic programming reported in this paper.

REFERENCES

[1] The Anti Phishing Working Group, http://www.antiphishing.org , Last accessed August 2012.

[2] Webopedia, http://www.webopedia.com/TERM/P/phishing.html.

[3] R. Feldman and J. Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, UK: Cambridge University Press,2007.

[4] K. J. Nishanth, V. Ravi, N. Ankaiah and I. Bose "Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts" *Expert Systems with Applications* ,vol. 39, no. 12, pp. 10583-10589, 2010.

[5] S. Abu-Nimeh, D. Nappa, X. Wang., and S. Nair, "A comparison of machine learning techniques for phishing detection", in *proceedings of the APWG ecrime researchers Summit*, Pittsburgh, USA, 2007.

[6] SpamBase, http://ftp.ics.uci.edu/pub/machine-learning-database/spambase/ , Last accessed: 2012.

[7] PhishingCorpus, http://monkey.org/~jose/wiki/doku.php?id=PhishingCorpus , Last accessed: June, 2012.

[8] V. S. Lakshmi and M. S. Vijaya, "Efficient prediction of phishing websites using supervised learning algorithms", *International Conference on Communication Technology and System Design*, Vol 30, 2012, pp. 798-805, 2011.

[9] PhishTank, http://www.phishtank.com , Last accessed: 2012.

[10] M. He, H. Shi-Jinn , P. Fan, M. K. Khan, R. Ray-Shine , L. Jui-Lin , C. Rong-Jian and S. Adi , "An efficient phishing webpage detector", *Expert systems with applications: An International Journal*, Vol. 38, Issue 10, 2011.

[11] R. Basnet, S. Mukkamala and A. H. Sung, "Detection of phishing attacks: A machine learning approach", *Studies in Fuzziness and Soft Computing*, vol. 226, pp. 373-383, 2008.

[12] SpamAssassin, http://www.spamassassin.apache.org , Last accessed: June, 2012.

[13] RapidMiner, http://www.rapid-i.com , Last accessed: May, 2012.

[14] H. Liu, J. Li and L.Wong, "A comparative study on feature selection and classification methods using Gene Expression Profiles and Proteomic Patterns", *Genome Inform*, ;vol. 13, pp. 51-60, 2002.

[15] A. Maher, M.A. Hossain, K. Dahal and T. Fadi, "Intelligent Fishing Detection System for e-Banking using fuzzy data mining", *Expert Systems with Applications*, vol. 37, no.12, pp. 7913-7921, 2010.

[16] A. Le, M. Athina and F. Michalis, "PhishDef: URL names say it all", *CoRR*, pp. 191-195, 2010.

[17] T. Fergus and C. Joe, "Phishing detection using Classifier Ensembles", *eCrime Researchers Summit*, Tacoma, WA, pp. 1-9, Sep-Oct. 2009.

[18] A. Maher, M.A. Hossain, T. Fadi and K. Dahal, "Intelligent phishing website detection using fuzzy Techniques", in *3rd International Conference Information and communication technologies: From theory to applications*, ICTTA, Damascus, Syria, pp. 1-6, April 2008.

[19] A. Ammar , W. Tat-Chee , A. Altyeb , M. Ahmad , A. Eman , M. Anbar, A. Esraa and S. Ramadass, "Evolving fuzzy neural network for phishing emails detection", *Journal of Computer Science,* vol. 8, no. 7, pp. 1099-1107.

[20] S. Afroz and R. Greenstadt, "PhishZoo: An automated web phishing detection approach based on profiling and fuzzy matching", *Technical Report DU-CS-09-03*, Department of Computer Science, Drexel University, Pennsylvania, USA, 2009.

[21] A. Herzberg and A. Gbara, "TrustBar: Protecting web users from spoofing and phishing attacks", Cryptology ePrint Archive: Report 2004/155, 2004.

[22] N. Chou, R. Ledesma, Y. Teraguchi and J.C. Mitchell, "Client side defense against Web based Identity Theft", *in preceedings of 11th annual network and distributed system Security Symposium (NDSS '04),* San Diego, CA, Feb. 2004.

[23] Waterken Inc., waterken YURL Trust Management for Humans, http://www.waterken.com/dev/YURL/Name/.

[24] NetCraft Anti-Phishing Tool, http://toolbar.netcraft.com/ Accessed: Nov 03, 2008.

[25] Google Safe Browsing Service in Mozilla Firefox Version 3, Accessed: Dec 01, 2008, http://code.google.com/apis/safebrowsing/firefox3_privacy.html.

[26] X. Chen, I. Bose, A. C. M. Leung and C. Guo, "Assessing the severity of phishing attacks: A hybrid data mining approach", *Decision Support System*, vol. 50, no. 4, pp. 662-672, Mar. 2011.

[27] V. Ravi, R. Lal, and N. R. Kiran, "Foreign exchange rate prediction using Computational Intelligence Methods", *International Journal of Computer Science and Industrial Management Applications*, ISSN 2150-7988, vol. 4, pp. 659-670, 2012.

[28] M. Chandrasekaran, K. Narayanan and S. Upadhyaya, "Phishing email detection based on structural properties" *NYS Cyber Security Conference,* Albany, New York, pp. 1-7, Jun. 2006.