CMPE-255 Significant Paper report

Shashank Raghuvanshi
shashank.raghuvanshi@sjsu.edu

Shreya Goyal
shreya.goyal@sjsu.edu

Rishitha Bandi
rishitha.bandi@sjsu.edu

Praveen Kumar
praveen.kumar@sjsu.edu

I. Introduction

For our significant paper we chose the research paper published in IEEE from *"2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)"* titled *"Detection of Phishing Emails using Data Mining Algorithms"* [1] by Sami Smadi, Nauman Aslam, Li Zhang, Rafe Alasem and M A Hossain.

The paper proposed a model for detecting phishing emails by focussing on the data preprocessing phase, to extract a set of features from different parts of emails. Phishing employs social engineering methods to gain access to user's sensitive information for malicious purposes. Phishing is usually done through mediums such as email, SMS, voice calls, and instant messages.

The authors emphasized the importance of detection of phishing in emails by referencing a report from *"Anti-Phishing Work Group (APWG) of 4th quarter of 2012"*, whose conclusion was that emails are the most common medium used for phishing, and further compared detection of phishing in emails with detection of phishing at the web page level. The authors argued the following points in favor of detecting phishing at the email level:

- Detecting phishing in emails should result in faster web load time as it will remove the necessity for analysing websites.
- The average lifespan of phishing websites is short (2.25 days) [2] so its also harder to analyse.
- Detecting phishing attempts at email level doesn't require user's attention, and can be more secure as users spend very less time to look for security indicators of websites, even when prompted [3].

II. Proposed Technique for phishing email detection

The authors mainly focussed on feature extraction from content and header of emails and then selected/created hybrid features from that which are used for classification.

A. Dataset

Authors used the 4559 legitimate emails from spam assasin [4] and 4559 phishing emails from nazario [5]. For preprocessing authors used mbx2eml tool to split the grouped dataset into multiple files and to convert them to eml format.

B. Feature Extraction

The authors extracted 23 hybrid features from emails, which are listed below:

- Sender domain name from email id matches the domain name of message ID (binary feature).
- Email is Html (binary feature)
- Email is multipart (binary feature)
- Email contains Html form (binary feature)
- Number of hyperlinks in email
- Number of different domains from hyperlinks
- Hyperlink target different from hyperlink text
- Number of hyperlinks that have domain different from sender domain
- Number of dots in hyperlink
- Number of urls that contain ipaddress
- Number of urls that contain hexadecimal
- Number of links that contain @ character
- Number of links in msg body containing non-standard port (instead of 80 and 443)
- Use of javascript (Binary feature)
- Number of images used as hyperlink
- Number of pictures with Image Map (images containing clickable areas with hyperlinks)
- Url containing Non-Ascii character
- Number of urls pointing to websites containing self signed certificate
- Message size in bytes
- Check if all domain names have DnS and Reverse DNS entry (binary feature false if numbers do not match)
- Text email (binary feature if content type is text/plain)
- Number of attachments
- Number of receivers

III. Performance Analysis

A. Experimental Setup

The authors divided the emails into 6 different sets of varying number of phishing and legitimate emails, manually selected to avoid duplication. The number of emails in each set were given as listed in TABLE I:

B. Evaluation Metrics

Following evaluation metrics were used by authors to compare the performance of algorithms:
M = Number of Phishing emails
D = Number of legitimate emails

TABLE I
SUMMARY OF DATASETS

Experiment	*No. of Emails*	*Phishing*	*Legitimate*
Exp1	1000	500	500
Exp2	2500	1250	1250
Exp3	4000	2000	2000
Exp4	5000	2500	2500
Exp5	7630	3828	3802
Exp6	9118	4559	4559

Nm = Number of emails correctly classified as phishing
Nd = Number of emails detected as legitimate
Nf = Number of legitimate emails detected as phishing
Np = Number of phishing emails detected as legitimate

1) **True Positive(TP):** Number of phishing emails correctly classified as such.

$$TP = \frac{Nm}{M}$$

2) **True Negative(TN):** Number of legitimate emails correctly classified as such.

$$TN = \frac{Nd}{D}$$

3) **False Positive(FP):** Number of legitimate emails incorrectly classified as phishing.

$$FP = \frac{Nf}{D}$$

4) **False Negative(FN):** Number of phishing emails incorrectly detected as legitimate.

$$FN = \frac{Np}{M}$$

Based on these 4, the following measures are derived:

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

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F - Measure = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

C. Results

Authors compared several algorithms for classification task and found Random Forest and J48 to be the best. The results in TABLE II are from the J48 algorithm used by the authors which is an implementation of C4.5 decision tree learner (based on ID3 algorithm).

D. Conclusions

The authors compared multiple previous works for the same task, and pointed out the limitations that they had, mostly related to either limited or unverified dataset, or omission of key evaluation metrics for the results. The Authors then showed that even with same algorithms, their approach had better performance on all key metrics on a verified dataset.

TABLE II
WEIGHTED AVERAGE CLASSIFICATION RESULT OF 6 EXPERIMENT WITH J48

Exp.	*Exp1*	*Exp2*	*Exp3*	*Exp4*	*Exp5*	*Exp6*
TPR	98.2%	97.60%	97.65%	98.00%	97.15%	96.75%
TNR	98.00%	99.28%	99.65%	99.32%	99.50%	99.47%
FPR	2.00%	0.72%	0.35%	0.68%	0.50%	0.53%
FNR	1.80%	2.40%	2.35%	2.00%	2.85%	3.25%
Accuracy	98.10%	98.44%	98.65%	98.66%	98.32%	98.11%
Precision	98.00%	99.27%	99.64%	99.31%	99.49%	99.46%
Sensitivity	98.20%	97.60%	97.65%	98.00%	97.15%	96.75%
F-Measure	98.10%	98.43%	98.64%	98.65%	98.31%	98.09%
AUC	98.20%	99.20%	98.60%	98.80%	98.50%	98.70%

TABLE III
AVERAGE CLASSIFICATION RESULTS FOR RANDOM FOREST, J48, PART, SIMPLECART AND MULTILAYER PERCEPTRON

Algorithm	*Random Forest*	*J48*	*PART*	*Simple CART*	*Multilayer Perceptron*
TP	98.22%	96.75%	96.47%	96.58%	94.91%
TN	99.52%	99.47%	99.74%	99.54%	97.96%
FP	0.48%	0.53%	0.26%	0.46%	2.04%
FN	1.78%	3.25%	3.53%	3.42%	5.09%
Accuracy	98.87%	98.11%	98.10%	98.06%	96.44%
Precision	99.51%	99.46%	99.73%	99.53%	97.90%
Sensitivity	98.22%	96.75%	96.47%	96.58%	94.91%
F-Measure	98.86%	98.09%	98.07%	98.03%	96.38%
ROC Area Avg.	98.90%	98.20%	98.50%	98.30%	98.30%

TABLE IV
AVERAGE CLASSIFICATION RESULTS FOR LIBSVM, BAYESNET, SMO, LOGISTIC REGRESSION AND NAIVE BAYES ALGORITHMS

Algorithm	*LibSVM*	*BayesNet*	*SMO*	*Logistic Regression*	*Naive Bayes*
TP	96.14%	94.06%	95.92%	94.80%	92.08%
TN	95.35%	97.19%	94.85%	95.72%	49.77%
FP	4.65%	2.81%	5.15%	4.28%	50.23%
FN	3.86%	5.94%	4.08%	5.20%	7.92%
Accuracy	95.74%	95.62%	95.38%	95.26%	70.93%
Precision	95.39%	97.10%	94.90%	95.68%	64.70%
Sensitivity	96.14%	94.06%	95.92%	94.80%	92.08%
F-Measure	96.14%	95.55%	95.41%	95.24%	76.00%
ROC Area Avg.	95.70%	98.40%	95.40%	98.10%	90.20%

IV. Key Learnings

Few of the key takeaway for us from the research paper are:

- Data preprocessing and feature selection are an important part of the data mining project, and can have significant impact on final model. We based our project on this research paper and used some of the hybrid features specified here.
- The authors seemed thorough in designing their experiments for performance evaluation, and in choosing the data set for the experiments.
- Random Forest algorithm seems to be most suitable for this task.
- Even simple features from emails can prove to be effective in detection of phishing emails.
- Since this method of detection does not rely on text analysis, so it also has no adverse impact from some of the techniques used by phishers to confuse text mining methods like hidden salting.

V. Suggested Improvements for future research

- The authors didn't use any features from text analysis, considering that there are many NLP based approaches that classify emails solely based on text analysis, so it seems, that analysis of text combined with mentioned features can potentially improve the effectiveness of the approach.
- Dataset used in research is old, so the effectiveness of model on recent techniques used by phishers can't be quantified.

References

[1] S. Smadi, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "Detection of phishing emails using data mining algorithms," in *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 2015, pp. 1–8.

[2] M. Aburrous, M. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7913 – 7921, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410003441

[3] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks," *International Journal of Human-Computer Studies*, vol. 82, pp. 69 – 82, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1071581915000993

[4] J. Mason, "Spam assasin," 2005. [Online]. Available: https://spamassassin.apache.org/publiccorpus/

[5] J. Nazario, "Nazario," 2007. [Online]. Available: http://monkey.org/jose/wiki/doku.php?id=phishingcorpus/