



**Efficient and Fast Server Based Phishing Detection
System Using URL Lexical Analysis**

by

**AMMAR YAHYA DAEFF
(1540211762)**

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computer Engineering

**School of Computer And Communication Engineering
UNIVERSITI MALAYSIA PERLIS**

2017

UNIVERSITI MALAYSIA PERLIS

DECLARATION OF THESIS

Author's full name : AMMAR YAHYA DAEF
Date of birth : 3-1-1984
Title : EFFICIENT AND FAST SERVER BASED PHISHING DETECTION SYSTEM
USING URL LEXICAL ANALYSIS
Academic Session : 2017-2018

I hereby declare that the thesis becomes the property of Universiti Malaysia Perlis (UniMAP) and to be placed at the library of UniMAP. This thesis is classified as :

- CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*
- RESTRICTED** (Contains restricted information as specified by the organization where research was done)*
- OPEN ACCESS** I agree that my thesis is to be made immediately available as hard copy or on-line open access (full text)

I, the author, give permission to the UniMAP to reproduce this thesis in whole or in part for the purpose of research or academic exchange only (except during a period of _____ years, if so requested above).


SIGNATURE

A5606042
(NEW IC NO. / PASSPORT NO.)

Date : 28/08/2017

Certified by:


SIGNATURE OF SUPERVISOR

Prof. Ir. Dr. R Badlishah Ahmad
NAME OF SUPERVISOR

Date : 28/08/2017

NOTES : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

ACKNOWLEDGEMENTS

First and foremost, "All the praises and thanks be to Allah, The Beneficent, The Merciful", "praise be to Him who has taught by the pen, who taught man that which he knew not"

First of all, my sincere thanks to Allah, who endowed me to complete this PhD thesis. I would like to thank my supervisor, Prof. Ir. Dr. R Badlishah Ahmad, for all your guidance, support, brilliant ideas, patience, and the opportunities you have presented me. Your managerial skills and uncompromising quest for excellence always motivated me to present the best of what I can be.

I would like to express my heartfelt gratitude to my supervisor Dr. Yasmin Yacob for all numerous hours of discussions, support and encouragement. She has been helpful, understanding and generous throughout the study. She has truly been a mentor and I owe her my deepest thanks.

Finally, I would like to express my deep gratitude and thank to my beloved family for their love, patience and support, especially my beloved mother, my beloved brothers and sister for their continuous support and supplication. In addition, I would like to thank my dearest friends in Iraq for their support and prayers.

TABLE OF CONTENTS

	PAGE
THESIS DECLARATION	i
ACKNOWLEDGEMENT	ii
TABLE OF CONTENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	x
LIST OF ABBREVIATIONS	xiii
LIST OF SYMBOLS	xvi
ABSTRAK	xvii
ABSTRACT	xviii
CHAPTER 1 INTRODUCTION	
1.1 Problem statement	5
1.2 Objectives	7
1.3 Scope of the Study	8
1.4 Research Significance and Contribution	8
1.5 Thesis Organization	9
CHAPTER 2 LITERATURE REVIEW	
2.4 Introduction	11
2.2 Overview of Phishing Attack's Vectors, Obstacles and Detection	11
2.2.1 Vectors for Phishing Attacks	11
2.2.2 Obstacles of Preventing Phishing	17
2.2.3 Taxonomy of Phishing Attack Detection Techniques	19

2.2.4	Technical Solutions to Date	19
2.2.4.1	Authentication	20
2.2.4.2	List Based Methods	25
2.2.4.3	The Whitelist Approach	26
2.2.4.4	Blacklist Method	28
2.2.4.5	Security Toolbars	32
2.2.4.6	Detection of Phishing Emails	35
2.2.4.7	Content Analysis of Web Pages	38
2.2.4.8	Analysis of URLs	45
2.3	Client-server based Phishing Prevention	50
2.3.1	Server Side Applications	50
2.3.2	Client Side Applications	51
2.4	Summary	58
CHAPTER 3 RESEARCH METHODOLOGY		
3.1	Introduction	60
3.2	Server Side Implementation (Phish Detect On Fly (PDOF))	60
3.3	The overall Research Methodology	61
3.4	Datasets	64
3.4.1	Phishing Data Collection	65
3.4.2	Legitimate Data Collection	66
3.4.3	Legitimate and Phishing Data Merging	67
3.5	Token based Classifier (TCL)	67
3.5.1	URLs Processing	68
3.5.2	TCL Statistical Operation	70

3.6	N-gram based Classifier (NGCL)	71
3.6.1	URLs Processing and Statistical Classifier	71
3.7	Language Model based Classifier (LMCL)	72
3.7.1	N-gram LM	83
3.8	ML based classifiers	77
3.8.1	LMCL as a Single Feature with Lexical Features for ML	78
3.8.2	Modified LM with lexical features for ML	81
3.8.2	Classification Algorithms	84
3.9	Classifiers Evaluation Metrics	88
3.10	Summary	90
CHAPTER 4 PRELIMINARY ANALYSIS FOR URL LEXICAL FEATURES		
4.1	Introduction	92
4.2	An Overview of The Experimental Setup	92
4.3	URL Tokens Discrimination Features Analysis	93
4.4	TCL Performance Evaluation	97
4.5	TCL Out of Sample Test	102
4.6	URL N-gram Analysis	103
4.7	NGCL Performance Evaluation	108
4.8	NGCL Out of Sample Test	112
4.9	URL Classification using LM Technique	113
4.10	LMCL Performance Evaluation	113
4.11	Comparison study	118
4.11	Summary	123

CHAPTER 5 MACHINE LEARNING CLASSIFIERS RESULTS

5.1	Introduction	125
5.2	An Overview of The Experimental Setup	126
5.3	URL Lexical Features Analysis	127
5.4	Classifiers Empirical Evaluation	130
5.5	J48 Testing using Mismatched URLs	135
5.6	Classifiers Performance Using Modified LM Features	136
5.7	J48 Testing using Mismatched URLs	141
5.8	Feature Ranking and Effectiveness	142
5.9	J48 Processing Time Analysis	143
5.9.1	Time for Feature Collection	144
5.9.2	Training and Testing Times	144
5.9.3	Processing Time Comparison	145
5.10	Summary	150

CHAPTER 6 CONCLUSION AND FUTURE WORK

6.1	Conclusion	152
6.2	Future Works	156

REFERENCES	158
-------------------	-----

LIST OF PUBLICATIONS	169
-----------------------------	-----

LIST OF TABLES

NO.		PAGE
2.1	Summary of URL based phishing detection techniques.	52
2.2	Summary of technical phishing prevention methods in term of scope, pros and cons.	55
2.3	The extent to which phishing protection methods meet requirements of a successful defence.	57
2.4	The advantages and disadvantages of phishing countermeasures applications.	58
3.1	LM data structure construction using 2-gram counts.	75
3.2	Length features employed for ML discrimination between phishing and legitimate URLs.	79
3.3	Counting features employed for ML discrimination between phishing and legitimate URLs.	80
3.4	Binary features employed for ML discrimination between phishing and legitimate URLs.	80
3.5	Description of modified LM features.	85
4.1	Maximum accuracy with Optimal threshold of 0.067 for all datasets.	101
4.2	TCL performance metrics on all datasets.	101
4.3	TCL overall error rates using mismatched datasets.	103
4.4	The percentage of overlapping 2-grams.	105
4.5	The percentage of overlapping 6-grams	105
4.6	Maximum accuracy with optimal threshold of 0.330 for all datasets.	111
4.7	NGCL performance metrics on all datasets.	111
4.8	NGCL overall error rates using mismatched datasets.	112
4.9	LMCL performance metrics on all datasets.	117

4.10	LMCL overall error rates using mismatched datasets.	118
4.11	Performance Metrics comparison of TCL, NGCL, and LMCL on TD.	122
4.12	Performance Metrics comparison of TCL, NGCL, and LMCL on TA.	122
4.13	Performance Metrics comparison of TCL, NGCL, and LMCL on OD.	123
4.14	Performance Metrics comparison of TCL, NGCL, and LMCL on OA.	123
5.1	Length features statistics of phishing and legitimate URLs.	128
5.2	Count features statistics of phishing and legitimate URLs.	129
5.3	Binary features statistics of phishing and legitimate URLs.	140
5.4	Classifiers results on TD.	134
5.5	Classifiers results on TA.	134
5.6	Classifiers results on OD.	135
5.7	Classifiers results on OA.	135
5.8	J48 error rates of training and testing using mismatched datasets.	136
5.9	J48 classification performance using modified 4-gram LM and URL lexical features on TD.	140
5.10	J48 classification performance using modified 4-gram LM and URL lexical features on TA.	140
5.11	J48 classification performance using modified 4-gram LM and URL lexical features on OD.	140
5.12	J48 classification performance using modified 4-gram LM and URL lexical features on OA.	140
5.13	Error rates of J48 using the modified 4-gram LM and URL lexical features under mismatched evaluation.	142
5.14	Features ranked by IGR on TD dataset.	143
5.15	J48 training and testing time analysis.	145

5.16	J48 training and testing time analysis using external features.	149
5.17	Accuracy and processing time comparison of ML classifiers using the proposed lexical features and external features.	150

©This item is protected by original copyright

LIST OF FIGURES

NO.		PAGE
1.1	APWG phishing website trends 1st quarter 2016.	2
2.1	Web, communication and malware attack vectors employed by phishers.	12
2.2	Safe browsing statistics relating to malware and phishing websites (Google, 2016).	17
2.3	A taxonomy of phishing attacks detection techniques.	20
2.4	The Firefox Netcraft Extension.	33
2.5	Classification of phishing countermeasures applications.	50
3.1	Overview of PDOF phishing URL detection system.	61
3.2	Overall research methodology block diagram of first part.	63
3.3	Overall research methodology block diagram of the second part.	64
3.4	Dataset merging methodology.	67
3.5	TCL training and testing flow.	69
3.6	NGCL training and testing flow.	72
3.7	Flowchart of N-gram extraction process on each URL.	76
3.8	LMCL training and testing flow.	77
3.9	Overview of ML based phishing URL detection system.	78
3.10	Flowcharts of feature extraction during training and testing phase.	82
4.1	An overview of the investigation towards analyzing URL lexical features for URL phishing detection.	92
4.2	Percentage of reused tokens in each dataset.	95

4.3	Overlap percentage of phishing datasets (a) and legitimate datasets (b).	96
4.4	Overlap percentage between phishing and legitimate datasets.	97
4.5	Optimum threshold selection for TD dataset.	99
4.6	Optimum threshold selection for TA dataset.	99
4.7	Optimum threshold selection for OD dataset.	100
4.8	Optimum threshold selection for OA dataset.	100
4.9	TCL classification results comparison on all datasets.	101
4.10	Percentage of reused grams using different n values in each dataset.	104
4.11	Percentage of reused 4-gram in each dataset.	106
4.12	4- gram overlap percentage of phishing datasets (a) and legitimate datasets (b).	107
4.13	4-gram overlap percentage between phishing and legitimate datasets.	107
4.14	Optimum threshold selection for TD dataset.	109
4.15	Optimum threshold selection for TA dataset.	109
4.16	Optimum threshold selection for OD dataset.	110
4.17	Optimum threshold selection for OA dataset.	110
4.18	NGCL results comparison on all datasets.	111
4.19	LMCL performance on TD dataset with different n values.	115
4.20	LMCL performance on TA dataset with different n values.	115
4.21	LMCL performance on OD dataset with different n values.	116
4.22	LMCL performance on OA dataset with different n values.	116
4.23	LMCL classification results comparison on all datasets.	117
4.24	TCL, NGCL, and LMCL error rates comparison on all datasets.	120

4.25	TCL, NGCL, and LMCL TPR comparison on all datasets.	120
4.26	TCL, NGCL, and LMCL TNR comparison on all datasets.	121
4.27	TCL, NGCL, and LMCL FPR comparison on all datasets.	121
4.28	TCL, NGCL, and LMCL FNR comparison on all datasets.	122
5.1	An overview of the investigation towards URL lexical based ML phishing detection.	126
5.2	Classification results comparison of J48, SVM and LR on TD dataset.	132
5.3	Classification results comparison of J48, SVM and LR on TA dataset.	133
5.4	Classification results comparison of J48, SVM and LR on OD dataset.	133
5.5	Classification results comparison of J48, SVM and LR on OA dataset.	134
5.6	Classification results comparison of J48, SVM and LR on TD dataset.	138
5.7	Classification results comparison of J48, SVM and LR on TA dataset.	138
5.8	Classification results comparison of J48, SVM and LR on OD dataset.	139
5.9	Classification results comparison of J48, SVM and LR on OA dataset.	139

LIST OF ABBREVIATIONS

AIWL	Automated Individual White List
AOL	America Online
APWG	Anti-Phishing Work Group
APWL	Anti-Phishing White List
ARFF	Attribute-Relation File Format
CCH	Contrast Context Histogram
DNS	Domain Name System
EMD	Earth Mover's Distance
FNR	False Negative Rate
FPR	False Positive Rate
HIPs	Human Interactive Proofs
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
LEO	Logo Extraction and Comparison
LM	Language Model
LMCL	Language Model based Classifier
LR	Logistic Regression
LUI	Login User Interface
MITB	Man in the Browser
ML	Machine Learning
MLE	Maximum Likelihood Estimation

NCL	National Consumers League
NGCL	N-gram based Classifier
NLP	Natural Language Processing
OA	Open Alexa
OCR	Optical Character Recognition
OD	Open DMOZ
PDOF	Phish Detect On Fly
PLSA	Probabilistic Latent Semantic Analysis
SC	Statistical Classifier
SIFT	Scale Invariant Feature Transform
SMS	Short Message Service
SPP	Single Password Protocol
SSL	Secure Sockets Layer
SVM	Support Vector Machine
TA	Tank Alexa
TCL	Token based Classifier
TD	Tank Alexa
TDF	Term Document Frequency
TIFF	Tagged Interchange File Format
UI	User Interface
URL	Uniform Resource Locator
US-CERT	The United States Computer Emergency Readiness Team

VoIP

Voice Over Internet Protocol

©This item is protected by original copyright

LIST OF SYMBOLS

$Tokenphishrate_i$	Token phish rate
$URLphishrate$	URL token phish rate
$info(T)$	The entropy function
C_j	URL class
T	set of choices
$Split(T)$	Information gain of split children
$K(x;x')$	SVM kernel function
$h(x)$	The distance to the boundary of decision
$P(w)$	4-gram probability

©This item is protected by original copyright

Sistem Pengesan Memancing Data yang Cepak dan Pantas Menggunakan Pelayan Berdasarkan Analisis Leksikal URL

ABSTRAK

Pengesanan serangan phishing ialah bidang penyelidikan yang signifikan untuk aplikasi keselamatan rangkaian. Laman web sah selalunya terdedah kepada serangan phishing. Phishing menyebabkan cabaran berterusan dan terus menjadi ancaman menerusi pelbagai vector seperti enjin carian, laman web palsu, emel dan mesej segera. Penipuan berbentuk ini telah berevolusi untuk kekal satu langkah ke hadapan oleh tindak balas terkini. Ia memanipulasi kelemahan pengguna yang menyebabkan penyelesaian masalah ini semestinya kompleks. Pengkelas phishing menggunakan ekstrak fitur untuk mengesan laman phishing dan ia bergantung kepada sama ada kandungan laman web, Pengesan Sumber Seragam (URL) atau kedua-duanya. Pengekstrakan fitur URL mengandungi hos dan maklumat leksikal. Di dalam tesis ini pengekstrakan fitur hanya berdasarkan fitur leksikal untuk mengurangkan kos pemprosesan disebabkan oleh pengekstrakan fitur maklumat hos. Fitur-fitur ini digunakan oleh pengkelas untuk mengesan laman web phishing. Kebanyakan strategi pengesanan serangan phishing melayan mekanisme pengesanan pelanggan. Di dalam tesis ini, teknik baru pengesanan serangan phishing di cadang untuk mencapai sistem yang pantas, tegap dan tepat dengan menggunakan fitur leksikal sahaja. Bahagian pertama tesis mempersembahkan analisa dan pembangunan untuk fitur leksikal URL sedia ada termasuk tokenisasi dan mekanisme n-gram yang mengekstrak dan menganalisa token dan pengagihan n-gram yang sah dan set data phishing diikuti dengan implementasi token berasaskan pengkelas (TCL) dan pengkelas berasaskan N-gram (NGCL). Oleh itu, TCL dan NGCL masing-masing memecahkan URL kepada token dan n-gram dan menggunakan pengagihan untuk proses klasifikasi. Juga, bahagian pertama tesis mencadangkan pengkelas berasaskan model bahasa (LMCL) yang membina model untuk kedua-dua kelas phishing dan sah untuk mengklasifikasi URL berdasarkan kemungkinan tertinggi dan dibandingkan dengan pengkelas TCL dan NGCL. Bahagian kedua tesis mencadangkan penggunaan output LMCL sebagai pengkelas fitur tunggal yang digabungkan dengan fitur leksikal URL untuk membina fitur keseluruhan yang digunakan oleh pengkelas Mesin Pembelajaran (ML). Kemudian cadangan untuk meminda output LMCL untuk mengekstrak model fitur sub-bahasa dan menggabungkan dengan fitur leksikal URL untuk melatih pengkelas ML. Berikutan strategi ML bernama J48, pengkelas Mesin Sokongan Vektor (SVM) dan regresi logistic (LR) digunakan untuk mengesan URL phishing. Prestasi penilaian telah dicapai di semua peringkat untuk memenuhi pengesanan serangan phishing yang pantas dan tepat. Sementara itu, kesemua pengkelas yang telah dicadangkan diuji menggunakan set data sebenar yang dikutip daripada pelbagai sumber untuk meneroka ketegaran teknik yang dicadangkan. Akhirnya, keputusan menunjukkan keboleharapan fitur leksikal berketepatan tinggi, tegar dan laju untuk pengesanan phishing URL. Diantara pengkelas yang dicadangkan, J48 dengan fitur cadangan menunjukkan keputusan keseluruhan terbaik dengan ketepatan 99% dan masa purata yang diperlukan untuk mengesan URL tunggal ialah 0.46 saat.

Efficient and Fast Server based Phishing Detection System Using URL Lexical Analysis

ABSTRACT

Phishing attack detection is a significant research area for network security applications. Legitimate websites is typically prone to phishing attacks. Phishing poses an ongoing challenge and continues to be a threat via numerous vectors such as search engines, fake websites, emails and instant messages. It has evolved its deceptions to remain one step ahead of the latest countermeasures. It exploits the weaknesses of the users which makes solving this problem especially complex. Phishing classifier uses the extracted features to detect the phishing websites and it depends on either the website's content, the Uniform Resource Locator (URL) or both of them. The URL feature extraction comprises host and lexical information. In this thesis, the feature extraction is based on the lexical features only in order to reduce the processing overhead due to the host information feature extraction. These features are utilized by a classifier to detect the phishing website. Most of the phishing attack detection strategies served the client side detection mechanisms. In this thesis, a new server side phishing attack detection technique is proposed to achieve fast, robust and accurate system by using lexical features alone. The first part of thesis presents analysis and development for the existing lexical features of URL including the tokenization and n-gram mechanisms which extract and analyze tokens and n-gram distribution of legitimate and phishing datasets followed by implementing Token based Classifier (TCL) and N-gram based Classifier (NGCL). Therefore, TCL and NGCL segment URLs into tokens and n-grams respectively and employ their distribution for classification process. Also, the first part of thesis proposing Language Model based Classifier (LMCL) which build a model for both of phishing and legitimate classes to classify URLs according to the highest probability and compared with TCL and NGCL classifiers. The second part of thesis proposing using the output of LMCL as a single classification feature in combination with URL lexical features in order to build the whole features that used by the Machine Learning (ML) classifiers. Then proposing to modify the output of LMCL to extract sub language model features and combined with URL lexical features to train ML classifiers. Regarding ML strategy J48, Support Vector Machine (SVM) and Logistic Regression (LR) classifiers are used for detecting the phishing URLs. The performance evaluation has been achieved regarding all these stages to meet a fast and accurate phishing attack detection. Meanwhile, all the proposed classifiers are tested using real life datasets collected from different sources in order to explore the robustness of the proposed techniques. Finally, the results showed a reliability of lexical features to provide high accuracy, robust and fast detection of phishing URLs. Among the proposed classifiers, J48 with the proposed features presents the best overall results with 99% accuracy and the time required to detect single URL is a 0.46 second on average.

CHAPTER 1

INTRODUCTION

The web has evolved widely in the life of people and since the beginning of Internet in 1990s, a lot of new security issues and threats appear continuously which constitute a challenge to users and security experts as well. Phishing is a cutting edge threat that has a deep impact on commercial and banking sectors by means of the Internet and delivers a huge misfortunes at the level of clients and organizations (Khonji, Iraqi, & Jones, 2013a). Phishing websites are highly similitude with the honest ones via trying to trap and bait users into these websites. Regarding this sort of attacks, phishers normally utilize technical and social designing traps together to begin their attacks. The social engineering attacks are focusing on users not on a system itself and intending to get the users information which are typically considered to be a touchy and confidential (Bozkir & Sezer, 2016).

Anti-Phishing Work Group (APWG)(Greg Aaron, 2016) reported that the number of phishing websites increased by 250% in the period from the last three months of 2015 to the first quarter of 2016 as shown in Fig. 1.1. The total number of discovered unique websites in the first quarter of 2016 is 289,371. Also, steadily rose per month was observed from October 2015 to March 2016 ranged from 48,114 to 123,555 respectively (Greg Aaron, 2016). These statistics demonstrate the significance to distinguish URLs and domain names to battle phishing. Additionally, the rise of online websites as highlighted in (Netcraft, 2016) reaches about one billion providing more services accessible on the Internet that can be targeted by phishers. Consequently, numerous new physical vectors, victims and targets get to be accessible letting space for a new sort of phishing

attacks to be executed.

Different attack vectors are used to launch phishing attack such as search engines, fake websites, advisement, email, instant message, social media or phone call (G. Liu, Qiu, & Wenyin, 2010). This assortment of phishing attacks leads to a difficult protection against this phenomenon and existing phishing detection methods just adapting to a few of them. In spite of the broad field of phishing attack vectors, a typical purpose of numerous vectors is the utilization of the link misleading victims for phishing websites. Utilization of obfuscated Uniform Resource Locator (URL) and domain names are widely used in phishing attacks (Aaron & Rasmussen, 2015). URL obfuscation lures users by misleading them to forged websites via a URL or website of a genuine website familiar to the victim (Aaron & Rasmussen, 2015; Cova, Kruegel, & Vigna, 2008).

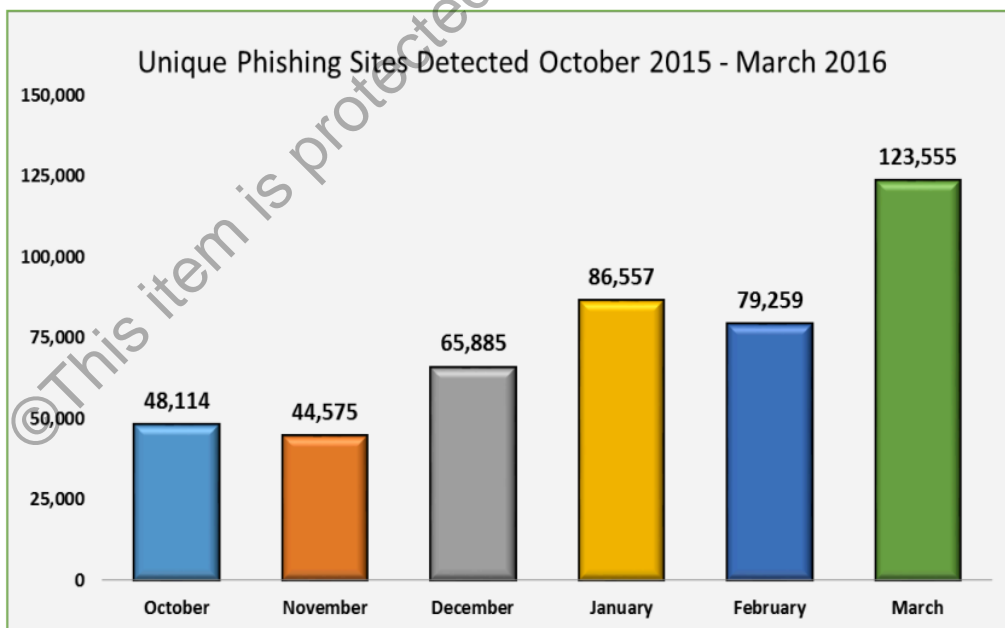


Figure 1.1: APWG phishing site trends 1st quarter 2016 (Greg Aaron, 2016).

Intelligent solutions based on phishing feature extraction (Abur-rous, Hossain, Dahal, & Thabtah, 2010; Almomani, Gupta, Atawneh, Meulenberg, & Almomani, 2013;

X. Chen, Bose, Leung, & Guo, 2011; Kazemian & Ahmed, 2015; Thomas, Grier, Ma, Paxson, & Song, 2011; Le, Markopoulou, & Faloutsos, 2011; J. Ma, Saul, Savage, & Voelker, 2011; Blum, Wardman, Solorio, & Warner, 2010a; J. Ma, Saul, Savage, & Voelker, 2009; Khonji, Iraqi, & Jones, 2011) depend on extracting important features of the website and after the extraction process these features utilized by an algorithm to decide or detect the phishing website. In general can be divided into content based and URL based solutions. Content based intercept and download the full contents of website for analyzing which can provide high detection accuracy with much more runtime overhead. In addition, it might accidentally provide more threats to users they look to keep safe from it. URL based techniques use a combination of host information and lexical features (Le et al., 2011; Thomas et al., 2011). Hosting information features need to be extracted from a remote server which in turn poses large latency to classify the URLs and prevent employing such methods for real time systems. While, URL lexical features are represented as bag-of-words result in huge vectors of features and cause processing overhead in addition to the low detection accuracy. Mostly, URL features are used to train a Machine Learning (ML) algorithms to generate a classifier to detect unseen URLs.

Generally, anti-phishing solutions can be positioned in different levels of attack flow where most researchers are focusing on client side solutions (Almomani et al., 2013; Khonji et al., 2013a; Heartfield & Loukas, 2016; Tewari, Jain, & Gupta, 2016; Aleroud & Zhou, 2017). The tools in client side include profile filter and browser toolbars. A few samples of a such tools can be specified by: CallingID¹, Spoof-Guard (Teraguchi &

¹<http://www.callingid.com/partners/safe-search/>

Mitchell, 2004), IE phishing filter², NetCraft³, CloudMark⁴ and eBay toolbar⁵. However, client side tools add more processing overhead which can lead to lose the trust and satisfaction of users. Other factor that has always been challenging for the researcher and security expert in browser based techniques is the mode to display the warning messages. Passive warning used to notify about phishing, such as change in colour, pop-up with textual information displayed at the corner or periphery of browser without interrupting browse activity is either unnoticed or neglected by Internet user (Wu, Miller, & Little, 2006; Aleroud & Zhou, 2017; Zeydan, Selamat, & Salleh, 2014).

On the other hand, server side solutions are usually based upon approaches which use content filtering and form the best means of defending against zero-hour or zero-day phishing attempts. For this reason, most new developments to address zero-day attacks are based on server side applications (Khonji, Iraqi, & Jones, 2012). The server side filters and classifiers applications based on machine-learning techniques for phishing attack detection are divided into sub-sections such as bag-of-words model (Blanzieri & Bryl, 2008; A. Hamid & Abawajy, 2011; Wardman, Stallings, Warner, & Skjellum, 2011), multi classifiers algorithms (Miyamoto, Hazeyama, & Kadobayashi, 2008; Islam & Abawajy, 2013), classifiers model based features (Islam & Abawajy, 2013; T.-C. Chen, Stepan, Dick, & Miller, 2014), clustering of phishing email (Bagirov, 2008; L. Ma, Yearwood, & Watters, 2009) and multi-layered system (Yearwood, Mammadov, & Banerjee, 2010; Olivo, Santin, & Oliveira, 2013; Abawajy & Kelarev, 2012). Generally, each has the same techniques, but has some differences in term of features extraction.

²<https://support.microsoft.com/en-us/kb/930168>

³<http://toolbar.netcraft.com/>

⁴<http://www.cloudmarkdesktop.com/>

⁵<http://pages.ebay.co.uk/help/accounttoolbar-install.html>

These previous studies of server side ML based techniques have built phishing classifiers to detect phishing websites using a combination of URL lexical features, hosting information, network traffic, and other strategies. Using lexical features only leads to low accuracy of detection which forces the designers of phishing classifiers to employ the other types of features such host information. Using such features required information to be looked up on a remote server. Though previous works had utilized URL lexical analysis as a component, what was lacking was the exploration of the full potential of a purely lexical approach to provide a high accurate and fast detection approach. Additionally, there is lack discussion of the delayed producing by these methods (A. Aggarwal, Rajadesingan, & Kumaraguru, 2012) and very little works stated the time required to detect a single URL which is considered unsuitable for real time application (Thomas et al., 2011; Le et al., 2011; Marchal, François, State, & Engel, 2014). As a consequence of restrictions in a current methods and the remembering that the most promising technique is URL analysis, especially the technique which depends only on lexical analysis and URL detection in a real time will be best familiar with minimum processing overhead. In addition, the main data entry points are usually a masqueraded URL (or link). Hence, this work proposing using URLs lexical features alone in order to explore the upper bound of performance can be achieved by URL lexical based phishing classifiers to provide high detection accuracy and minimum processing time to classify a single URL.

1.1 Problem Statement

Server side phishing detection systems are considered as the ideal solution to detect zero-day attacks online (Khonji et al., 2012; Almomani et al., 2013; Thomas et al., 2011). These systems should be lightweight enough to support the real time process and