PHISHING WEBSITE DETECTION

NUR SHOLIHAH BINTI ZAINI

Bachelor of Computer Science
(Computer System and Networking)

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : _____

Date of Birth : _____

Title : _____

_____

_____

Academic Session : _____

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☑	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang
2. The Library of Universiti Malaysia Pahang has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

_____	_____
(Student's Signature)	(Supervisor's Signature)
_____	_____
New IC/Passport Number	Name of Supervisor
Date:	Date:

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Perpustakaan Universiti Malaysia Pahang,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak,
26300, Gambang, Kuantan.

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name
Thesis Title

Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,

_____
(Supervisor's Signature)

Date:

Stamp:

Note: This letter should be written by the supervisor, addressed to the Librarian, *Perpustakaan Universiti Malaysia Pahang* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Computer Science in Networking.

_____

(Supervisor's Signature)

Full Name : DR MOHD FAIZAL BIN AB RAZAK

Position : SENIOR LECTURER

Date : 7 JANUARY 2019

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

_____

(Student's Signature)

Full Name : NUR SHOLIHAH BINTI ZAINI

ID Number : CA15080

Date : 7 JANUARY 2019

PHISHING WEBSITE DETECTION

NUR SHOLIHAH BINTI ZAINI

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science

Faculty of Computer System and Software Engineering

UNIVERSITI MALAYSIA PAHANG

DECEMBER 2018

ACKNOWLEDGEMENTS

In the accomplishment of this research successfully, many people have best owned upon me their blessings and the heart pledged support, this time I am utilizing to thank all the people who have been concerned about this research.

Primarily I would to thank Allah for being able to complete this research with success. Then I would like to thank my supervisor, Dr Abdul Ghani Ali Ahmed and Dr Mohd Faizal bin Ab Razak whose valuable guidance has been the ones that helped me patch this research and make it success. Their suggestion and instruction have served as major contributors towards the completion of the research.

Then I would like to thank my parent, Zaini bin Taib and Noridah binti Ismail for giving me support both morally and finance. Next, I would like to express my gratitude to my friends, Ibrahim bin Othman, Nurul Farah Anisah binti Jenimen, Ong Vienna Lee and my roomates who have been helped me with their valuable suggestion and moral support in various phases of the completion of the research.

Last but not least I would like to thank the entire people who directly or indirectly contribute on completing this research.

ABSTRAK

Internet telah menjadi sebahagian daripada aktiviti sosial dan kewangan harian kami. Internet adalah penting bukan sahaja untuk pengguna individu, tetapi juga untuk organisasi, lebih lebih lagi sebagai organisasi yang menawarkan perdagangan dalam talian dapat memperoleh kelebihan daya saing dengan menawarkan pelbagai perkhidmatan kepada pelanggan global. Internet memungkinkan untuk mencapai pelanggan di seluruh dunia tanpa sekatan pasaran dan dengan e-dagang yang berkesan. Akibatnya, bilangan pelanggan yang menggunakan Internet untuk membuat pembelian mereka meningkat dengan ketara. Beratus-ratus juta dolar dipindahkan setiap hari melalui internet. Jumlah wang ini menarik perhatian penjenayah siber untuk menjalankan aktiviti haram mereka. Oleh itu, pengguna Internet mungkin terdedah kepada pelbagai jenis ancaman web yang boleh menyebabkan kerugian kewangan, penipuan kad kredit, kehilangan data peribadi, menjejaskan reputasi organisasi dan kehilangan kepercayaan dalam perkhidmatan e-dagang dan perbankan dalam talian oleh pelanggan. Oleh itu, kesesuaian internet untuk urus niaga komersial akan dipersoalkan. Phishing dianggap sebagai ancaman web yang didefinisikan sebagai seni penyamaran sebagai laman web sebenar untuk mendapatkan nama pengguna, kata laluan dan butiran kad kredit. Dalam kajian ini, fenomena Phishing akan dibincangkan secara terperinci. Di samping itu, kami membentangkan kajian mengenai cara penyelidikan berkenaan topik ini. Tambahan pula, penyelidikan ini bertujuan untuk mengenal pasti perkembangan terkini dalam phishing dan langkah berjaga-jaga, serta menjalankan kajian dan penilaian komprehensif terhadap penyelidikan ini untuk menutup jurang yang masih wujud dalam topic ini. Penyelidikan ini tertumpu terutamanya pada kaedah pengesanan phishing data berasaskan web, tidak tertumpu kepada kaedah pengesanan berasaskan e-mel.

ABSTRACT

The Internet has become an integral part of our daily social and financial activities. The Internet is important not only for individual users, but also for organizations, as organizations that offer online commerce can gain a competitive advantage by serving global customers. The Internet makes it possible to reach customers all over the world without market restrictions and with effective e-commerce. As a result, the number of customers using the Internet to make their purchases is increasing significantly. Hundreds of millions of dollars are transferred every day over the internet. This amount of money tempted the fraudsters to carry out their illegal activities. Therefore, Internet users may be vulnerable to various types of web threats that may cause financial harm, credit card fraud, loss of personal data, potential damage to brand reputation, loss of trust in e - commerce and online banking by customers. Hence, the suitability of the internet for commercial transactions is questionable. Phishing is considered a form of web threats that is defined as the art of impersonating a legit website to obtain usernames, passwords and credit card details. In this study, the phishing phenomena will be discussed in detail. In addition, we present a study on the state of research on the topic. Furthermore, we aim to identify the current developments in phishing and its precautionary measures, and to conduct a comprehensive study and evaluation of this research to close the gap that still exists in this area. This research focuses primarily on web - based phishing detection methods, not email - based detection methods.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
BART	Bayesian Additive Regression Trees
CART	Classification and Regression Trees
DNS	Domain Name System
FN	False Negatif
FP	False Positive
FPR	False Positive Rate
GNU	General Public License
GUI	Graphic User Interface
HTML	Hypertext Markup Language
IP	Intenet Protocol
IT	Information Technology
KNN	K-Nearest Neighbors
LR	Logistic Regression
MLBDM	Machine Learning Based Detection Method
MLP	Multi-Layer Perceptron
N	Unknown
NB	Naive Bayes
NN	Neural Networks
PD	Phishing Detection
PW	Phishing Website
PWD	Phishing Website Detection
Q1	First Quarter
Q3	Third Quarter
Q4	Fourth Quarter
RF	Random Forests
ROC	Receiver Operating Characteristics
SDLC	Software Development Life Cycle
SFH	Server Form Handler
SSL	Secure Sockets Layer
SVM	Support Vector Machines

TF–IDF	Term Frequency Inverse Document Frequency
URL	Uniform Resource Locator
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1 Background Overview

Phishing defined as a way of attempting to acquire information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication. It is a tool used by cyber criminals to steal personal information from user. The criminals will create a fake websites that look the same as the real websites.

User will get fraud by entering their confidential information such as password, bank details and account credentials into the fake websites. The fake website usually provides an embedded link to confirm the account details of the user. The criminal will then use the information provided to access the account to buy stuff, transfer money, or other damaging activities.

Phishing fraud has become the biggest threat to Internet security, according to "Chinese Network Security Report in the first half of 2011" issued by 360 SafetM, the largest security company in China. The number of phishing attacks has increased significantly in recent years, as reported by International Anti-phishing Alliance. It has become particularly urgent to find effective phishing detection methods.

1.2 Problem Statement

Internet is very useful and beneficial for everyone. The activities become online, for example, online shopping, online banking, online communication and cloud storage. However, this service is unfortunately not secure due to phishing websites.

Although there are many existing system for detecting phishing website, this systems are still unable to detect and prevent all kinds of phishing.

Moreover, existing system still have very high false alarm rated in differentiating between the phishing and normal website.

1.3 Project Objective

The main objective of this project is to detect the phishing websites. The general objectives to achieve for develop system:

i. To investigate security flaws by analyzing the state-of -the-art phishing detection system
ii. To propose a phishing detection system that analyzes website applications using machine learning
iii. To evaluate the proposed system in terms of accuracy of detection

1.4 Project Scope

For this project, it can be categorize into three scopes which are:

i. Platform
    - This application can be run on websites.
ii. Functionality
    - Computer user can be able to detect the phishing websites.

iii. User

- Every computer users (students, finance department and Government workers)

1.5 Significance

From this study, this research will find out the benefit of detecting phishing websites. There benefits that will receive are:

i. Provides organizations the safety of their websites

ii. Give banking institutions' official website more secure.

iii. Prevents internet user from get trick and have financial loss.

1.6 Thesis Content

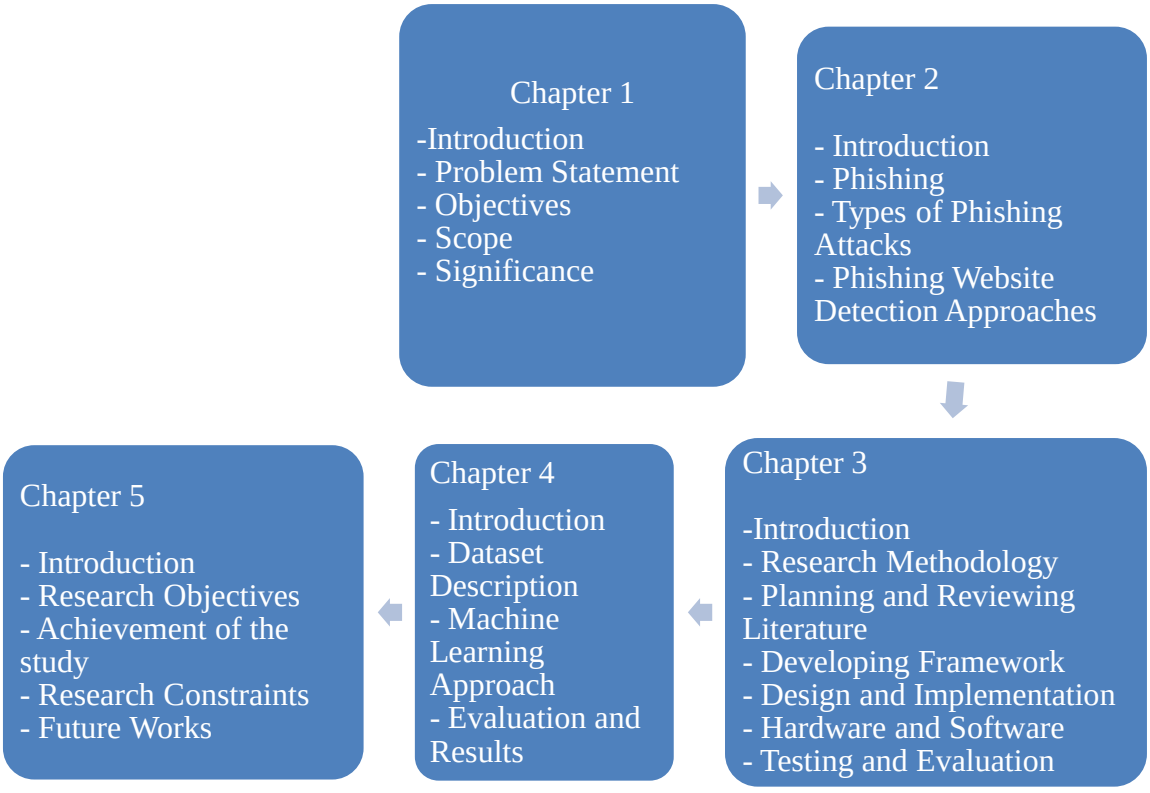


Figure 1 Summary of Each Chapter

This research will include five chapters. In Chapter one this research will discuss about introduction on this system, describe briefly information that

relate with current issue. It is also come out with problem statement and solution for this project.

In Chapter two, it is about the literature review. It describes the meaning of phishing and types of phishing attacks. This chapter also elaborate three different types of phishing attack detection approaches.

For Chapter three, it will discuss about methodology that for development and justification using the chosen methodology. In this chapter also will explain phase for development, software and hardware use for this system. Most important things are the implementation of development of the project and testing method that explain at this chapter.

Chapter four is about testing method use in project and result of outcome from the project. It also contains user manual and other attached in appendix. The consequence of the project must be match and achieved with objective of the project.

For last chapter in this stuy, to conclude overall of the project based on the output that fit with objective, implementation of methodology use for project that need specify software and hardware along with constraint system and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This research has been discussed the introduction of research in Chapter 1, which consists of the problem statement, objective, significance and scope. In this chapter, this study will discuss the relevant literature review to understand the system technique and how the PW (Phishing Website) can be detected. Therefore, existing PW detection works will be further developed in order to justify the current work.

2.2 Phishing

Phishing tries to obtain sensitive data such as usernames, passwords and credit card details, often for malicious reasons, by disguising an electronic communication as a trustworthy entity.

Email spoofing or instant messaging typically carries out phishing, and it often directs users to enter personal information at a fake website, the look and feel identical to the legitimate one and the only difference is the URL (Uniform Resource Locator) of the website in concern. Communications platform such as social web sites, auction sites, banks, online payment processors or IT (Information Technology) administrators often are used to lure victims. Phishing emails may contain links to websites that distribute malware.

Over the years, phishing attacks have increased globally. The total number of phishes detected was 263,538 in Q1(First Quarter) 2018. This increased by 46 percent compared to the 180,577 observed in Q4 (Fourth Quarter) 2017. It was also considerably more than in Q3 (Third Quarter) 2017 in 190,942 (APWG, 2018). This is shown at Figure 2.1 below.

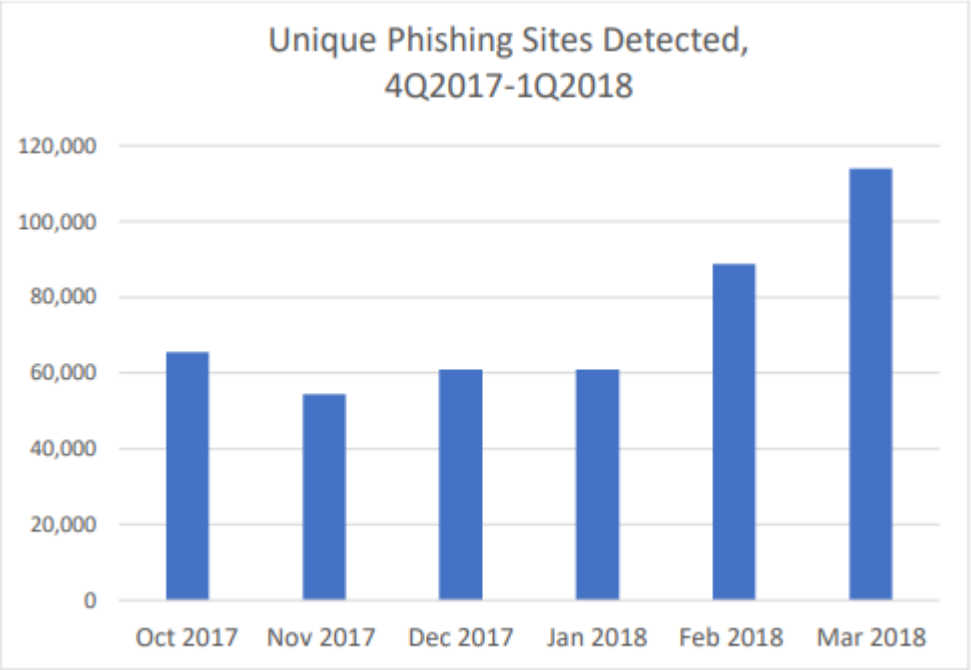


Figure 2Unique Phishing Sites Detected

2.3 Type of Phishing Attacks

There are currently various types of phishing attacks. It has been categorized into three different types as below:

2.3.1 Deceptive Phishing

Deceptive phishing is the messages required to verify account information, requesting that users re-enter their information, bogus account charges, unwanted account changes, new free services requiring immediate action, and many other malicious sites are sent to many recipients in the hope that the unsuspecting person will react by clicking on a link to or signing on a fake site.

2.3.2 Malware-based Phishing

It refers to attacks that lead to the installation and execution of malicious software on computers of users. Malware is generally introduced as an email attachment that can be downloaded. Malware commonly installed in phishing attacks includes key loggers and screen grabbers, spyware that captures and logs input keyboards or display the screen and sends information to the phisher. In other cases, the

target of the attack is to control the computer of the victim (Chaudhry, Chaudhry, & Rittenhouse, 2016).

2.3.3 Content-Injection Phishing

The injection of content is a technique in which the phisher changes a part of the content on a reliable website page. This is done in order to mislead the user to go to a page outside the legitimate website where personal information is to be entered (Nisha & Madheswari, 2016).

2.4 Phishing Website Detection Approaches

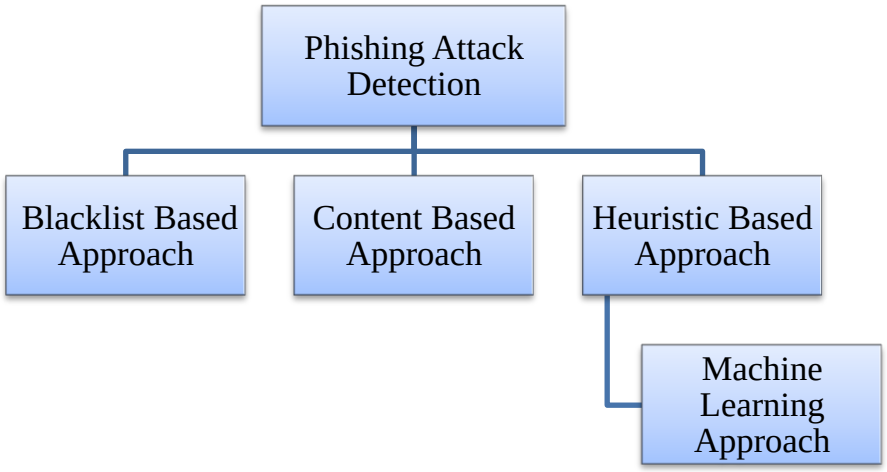


Figure 3 Phishing Website Detection Approaches

2.4.1 Blacklist-based Approach

A blacklist is a list of malicious URLs. Blacklist is obtained using a number of methods, such as heuristics from web crawlers, manual voting and honeypots. When a website is visited, the browser refers it to the blacklist to check whether the current URL is included in the list. If so, it indicates that it is a malicious website and therefore the browser warns users not to submit sensitive information.

The drawback of this approach is that blacklists cannot normally cover all phishing websites because a newly created fraudulent website takes a considerable amount of time before it is added. This time gap between launching the suspicious website and adding it to the list may be sufficient for phishers to accomplish their goals.

The detection process should therefore be extremely fast, usually when the phishing website is uploaded and the user begins to submit his credentials.

If the blacklist update process is slow, website phishers will be able to carry out attacks without being added to the blacklist. Blacklists are updated at different speeds, and in a study, scientists estimated that approximately 47 percent –83 percent of phishing URLs are displayed on blacklists nearly 12 hours after launching. The same study found that zero hours of defense from the most well-known toolbars on blacklists claimed a TP rate of 15 percent – 40 percent (Sheng et al., 2009). Therefore, an efficient blacklist needs to be updated immediately in order to keep users safe from phishing.

Netcraft is a small package of software that is activated when a user browses the Internet (Netcraft Ltd., 2008). Netcraft is based on a blacklist of fraudulent websites recognized by Netcraft and the URLs submitted an by the users and verified by Netcraft. Netcraft displays the server location where a web page is hosted and this is especially useful for users who are familiar with it. The main features used by Netcraft for calculating the risk rate for each site and deciding whether to add it to the blacklist are:

i. How old is the domain name where the website is located.

ii. Domain names not included in the Netcraft database.

iii. The existence of any previously hosted phishing web pages in the same domain.

iv. Use IP addresses or hostnames in the URL.

v. Country history and Internet service provider regarding the hosting of phishing websites

vi. The history of top- level phishing web sites domains.

vii. How prominent is the website in the Netcraft Toolbar community.

The main problem with Netcraft is that the final decisions about the legitimacy of the website are made primarily by the Netcraft server and not by the user's computer.

Therefore, if for any reason the connection to the server is lost, the user is under threat and vulnerable during this period.

2.4.2 Content-based Approach

The proposed method suggests using CANTINA, a content- based technique for the detection of phishing websites using the term- frequency– inverse document-frequency (TF–IDF) measurements (Xiang, Hong, Rose, & Cranor, 2011).

CANTINA then examines the content of the webpage to determine whether it is phishing or not using TF– IDF. By counting its frequency, TF– IDF produces weights that assess the importance of the word to a document.

CANTINA operates as follows:

i. For a given web page, calculate the TF– IDF.

ii. Take the top five TF– IDF terms and add them to the URL in order to find the lexical signature.

iii. Put the lexical signature in a search engine.

If the result of the search for N tops contains the current website, it is considered a legitimate website. However, if not, it is a phishing website. In experiments, N was set to 30. If the search engine returns zero results, however, the website is labelled as phishing. This argument was the main disadvantage of using such a technique, as this would increase the false positive rate (FP).

CANTINA identifies the phishing website successfully, but disables the extraction of the keywords. Some attackers are now using HTML hidden text to avoid the keyword extraction technique. In addition, CANTINA suffers from a performance challenge because it takes a considerable amount of time to query Google.

2.4.3 Heuristic-based Approach

The third technique is known as heuristic approaches, which collect some features from the website to identify them as either phishing or legitimate. Unlike the blacklist method, a heuristic solution can identify in real time newly created phishing

websites. The efficiency of the heuristic methods depends on the selection of a set of discriminative features that could help to distinguish the website type. The heuristic approach uses the HTML or URL signature that identifies the phishing web pages. Several studies are carried out based on this approach.

SpoofGuard is one of the solutions that employ heuristics. It is a plug-in anti-phishing browser (N. Chou, Ledesma, Teraguchi, Mitchell, & Ca, 2004). This approach uses a combination of stateless page evaluation, full page evaluation and outgoing post data examination to calculate the spoof value. If the spoof index is larger than a pre - defined threshold value, the page will be classified as a phishing page and the user will be notified of this page.

2.4.3.1 Machine Learning Approach

In addition to the above mentioned techniques, a set of literature is intended to evaluate the performance of machine learning and data mining algorithms.

The authors compare the predictive accuracy of a number of machine learning methods, such as Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Random Forests (RF) and Neural Networks (NN). A dataset consists of 1171 phishing emails and 1718 legitimate emails in the comparative experiments were used. To learn and test the classifiers, a set of 43 functions was used. Experiments show that RF has the lowest error rate of 7.72% followed by CART of 08.13%, followed by LR of 08.58%, followed by BART of 09.69%, then Support Vector Machines (SVM) of 09.90%, then NN of 10.73%. The results show, however, that no optimum classifier can be used to predict phishing sites.

(Miyamoto, Hazeyama, & Kadobayashi, 2009) evaluate the performance of machine learning based detection methods (MLBDMs) including AdaBoost, Bagging, SVM, CART, LR, RF, NN, Naive Bayes (NB) and BART. A dataset consist of 1500 phishing websites and 1500 legitimate websites were used in the experiments. The evaluation based on 8 heuristics presented in CANTINA .

Before starting their experiments a set of decision were made by the authors as follow:

i. The number of trees in Random Forest is set to 300.

ii. For all experiments need to be analysed iteratively the number of iteration was set to 500.

iii. Threshold value was set to 0 for some machine learning techniques such as BART.

iv. Radial based function was used in support vector machine.

v. The number of hidden neurons was set to 5 in the neural network experiments.

The experiments showed that 7 out of 9 MLBDMs outperform CANTINA's accuracy and those are: AdaBoost, Bagging, LR, RF, NN, NB and BART.

In (Abu-Nimeh, Nappa, Wang, & Nair, 2007) the authors compare the predictive accuracy of a number of machine learning methods those are LR, CART, BART, SVM, RF, and NN. A dataset consist of 1171 phishing emails and 1718 legitimate emails were employed in the comparative experiments. A set of 43 features were used to learn and test the classifiers. The experiments show that RF has the lowest error rate of 7.72%, followed by CART 08.13%, followed by LR 08.58%, followed by BART 09.69%, then SVM 09.90%, and finally NN with 10.73%. However, the results indicate that there is no optimal classifier might be used to predict phishing websites. For instance, the FP rate when using NN is 5.85% and the false negative (FN) rate is 21.72% whereas the FP rate for RF is 8.29%, and the FN rate is 11.12%, which means that NN outperform RF in term of FN but RF outperform NN in term of FP.

2.4.4 Comparison between Phishing Website Detection Approaches

Table 2.1 Comparison between Phishing Website Detection Approaches

Types	Feature	Limitation
Blacklist-based Approach	-A blacklist is a list of malicious URLs. -When a website is visited, the browser refers it to the blacklist to check whether the current URL is included in the list.	- Blacklists cannot normally cover all phishing websites because a newly created fraudulent website takes a considerable amount of time before it is added.
Content-based Approach	- Examines the content of the webpage to determine whether it is phishing or not.	- Disables the extraction of the keywords.
Heuristic-based Approach	-Collect some features from the website to identify them as either phishing or legitimate	- Depends on the selection of a set of discriminative features that could help to distinguish the website type.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In the continuation on previous chapter which is the literature review, the elaboration of the definition of phishing website and the details about it provide the understanding of to proceed to this chapter. The previous techniques used to detect the phishing website as well as the previous researches have been made on detecting phishing website. In this chapter, details explanation will be given on the techniques and method as well as the features that I have chosen for this project.

There are lots of methodologies that can be defined. But for this research, we will focus on software development life cycle (SDLC). This methodology consist of 5 phases which are a planning phase, analysis phase, design phase, an implementation phase and maintenance phase. In SDLC, the final analysis is the most important factor for the success of a project may be how closely the plan was followed. The documentation is important regardless of the type of model selected or developed for each application and is usually executed in parallel with the development process. As mentioned in the last section of previous chapter, the technique that I will be using in this project is heuristic-based detection by using URL. Among the techniques which were discussed in previous chapter, the model chosen for this research project is waterfall methodology.

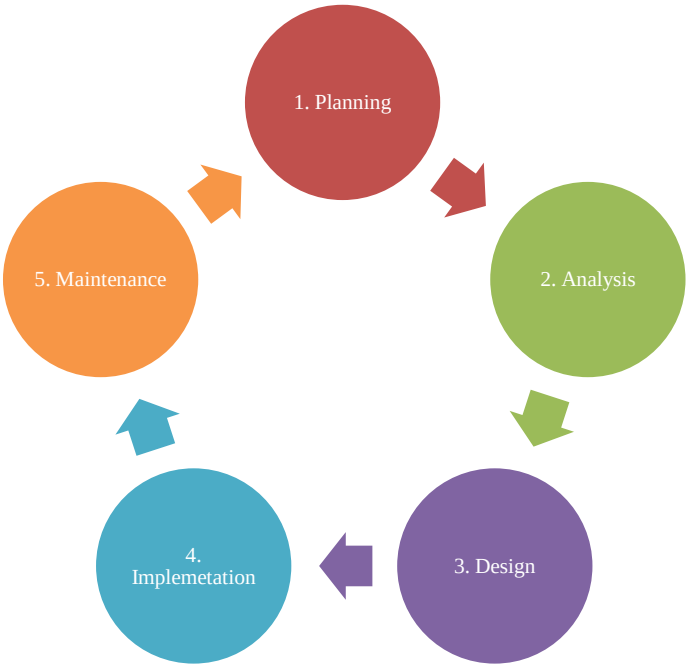


Figure 4 Software Development Life Cycle (SDLC)

3.2 Research Methodology

This research methodology consists of four main phases: literature study, development of a new framework, design and implementation and testing and evaluation. This methodology is suitable and adapted for use in this research project, as the stages can be reviewed and updated to get the best results. This research methodology differs from other system development lifecycles because its approach is more focused on conducting a detailed and careful research of the research topic.

The first stage of this research methodology is the review of the literature. Existing studies on the research topic will be carefully reviewed and analyzed at this stage. Then the research definitions are characterised as objectives and problems statement. The next step is the development of the frame. In this phase, the critical analysis of existing studies will be considered in the selection of a suitable method and algorithm to be used in this research. Now that the framework of the research project has been developed, the next stage of the design and implementation of the research will take place. The technical requirements such as language, hardware and software are specified for this stage to set up the research experiment. When the research

requirements are designed and prepared, the actual implementation of the research project is applying to design the detection model. Once the implementation has been completed, the research experiment is tested and evaluated to determine the limitations of research and the improvements that can be made in future research.

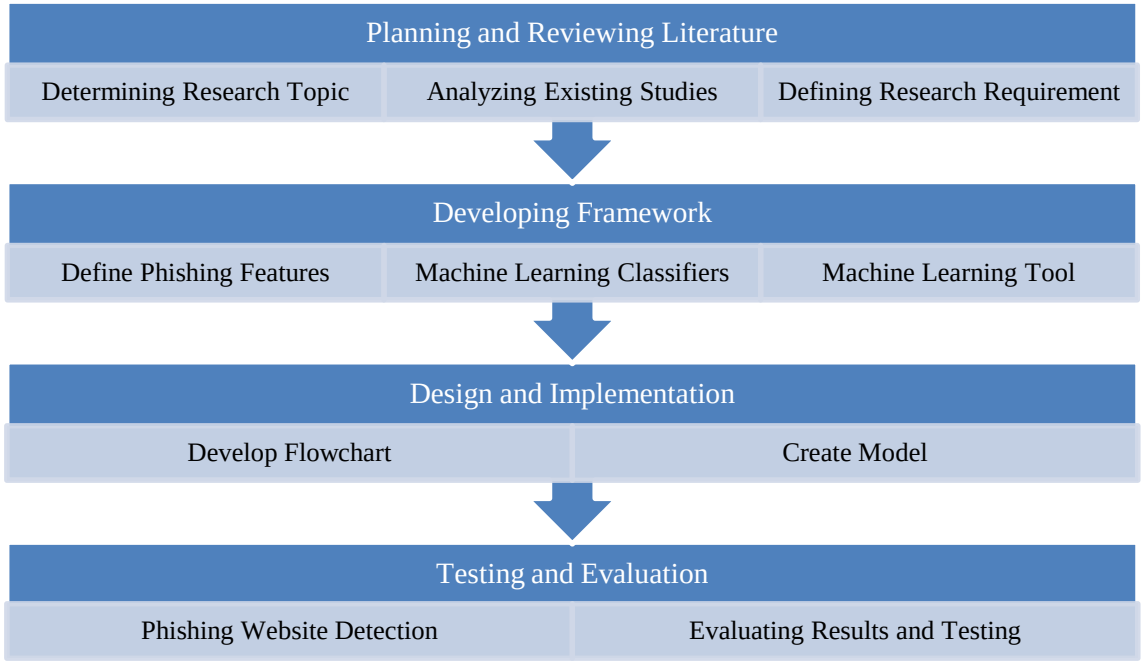


Figure 5 Main Stages for Research Methodology

This research methodology is adapted in this research project because it can be returned to previous stages with minimal losses in order to implement new research improvements. Not only that, this methodology allows changes to any stage to be made from time to time to solve problems during the current stage. Finally, this research methodology gives researchers the advantage of adapting easily to the needs of the research project.

3.3 Planning and Reviewing Literature

The research methodology's primary phase is research planning and literature review related to the research topic. Before examining the existing studies, the conceptualization is completed to determine the type of the relevant research topic. When the topic of research is chosen, related journals, articles and studies are collected to be studied. The study of existing studies enables to understand the research topic. This allows the problem statement, the objective and the scope of this research to be defined.

The resources we have collected are via Internet journals, previous student references and online e - books. The existing scheme studies are carefully analyzed and filtered in accordance with the relevance of the research topic. The collected information should be relevant for research, so that it can be used in the development of this research.

Based on gathered information, the different methods and techniques are learned to identify which type of method and technique is the best to solve the problem on the website application, especially phishing. Since their security problem is the main concern in the application of the website, we focus on the phishing detection (PD) for the website in this research. Existing PD research works are critically analyzed and classified depending on where the phishing code mechanism is performed. Each PD scheme proposed is analyzed to determine its contributions and limitations. This information is vital to determine the methodology used by the researches to perform their experimental tests. Therefore, research limitations will be avoided in this study.

3.4 Developing Framework

Based on our study of the existing phishing code detection scheme, we decided to develop a phishing code detection scheme that uses a machine learning approach and detects the features. Figure 8 illustrates the development of a PWD framework.

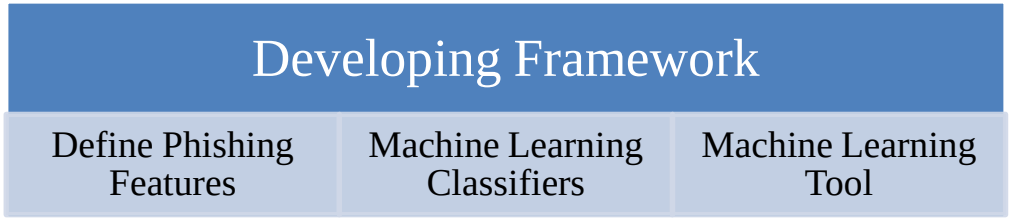


Figure 6 Development of PWD Framework

3.4.1 Define Phishing Features

In academic literature and commercial products there are many algorithms and a wide range of data types for phishing detection. A phishing URL and the page have several features that can be distinguished from a malicious URL. An attacker can, for example, register long and confusing domains to hide the real domain name.

In this study, we will investigate the URL based features. The URL is the first thing to analyze a website that decides whether or not to phish. Phishing domain URLs has some distinctive features. Features associated with these points are obtained with the processing of the URL. The URL-Based Features that will be going to investigate are given below:

i. Address Bar based Features

ii. Abnormal Based Features

iii. HTML and JavaScript based Features

iv. Domain based Features

3.4.2 Machine Learning Classifiers

Machine learning is a type of artificial intelligence (AI) that can learn without explicit programming. It is also capable of predicting future decisions and improving

decisions when exposed to new data. The prediction process is based on the search through the data set that looks for patterns. This is also referred to as learning. The learning process and prediction results depend on the classifier types. This technique was widely used to classify samples particularly in the area of intrusion detection systems (phishing and normal). The two common types of machine learning are supervised and unsupervised machine learning. The prediction process is based on the search through the data set that looks for patterns. This is also referred to as learning. The learning process and prediction results depend on the classifier types. This technique was widely used to classify samples particularly in the area of intrusion detection systems (malware and normal). The two common types of machine learning are supervised and not supervised.

This research applies the supervised machine learning approach, since the sample data set have labels (phishing and normal). In addition, supervised machine learning offers good results through the reduction of errors. This study implement four classifiers in order to observe the distinctive results noted in the various machine learning classifiers. The four classifiers are: Random Forest (RF), J48, Multi-Layer Perceptron (MLP) and K-Nearest Neighbors (KNN). They are explained as below:

Random Forest (RF): Random Forest RF is a well known method of collective learning for supervised classification or regression. This machine learning technique works by building a random set of decision trees during training and producing the class which is the class mode (classification) or mean prediction (regression) of the individual trees (Vanhoenshoven, Gonzalo, Falcon, Vanhoof, & Mario, 2016).

MLP: Multi - layer perceptron is a model for the artificial neural network. MLP consists of multiple node layers interacting through weighted connections (Amalina, Ali, Badrul, & Abdullah, 2016).

J48: J48 is an ID3 extension. The additional features of J48 account for missing values, the pruning of decision trees, continuous value ranges and rules derivation. J48 is an open source Java in the WEKA data mining tool Realization of the C4.5 algorithm.

KNN: KNN is one of the simple classifiers for machine learning that works well in classifications. It is a lazy classifier type of learning. The classifier uses training

samples to predict the label, whereas by labelling the sample the user defines and classifies the KNN.

3.4.3 Machine Learning Tool

Functionality of machine learning tools for data analysis that automates the development of the analysis model. This model allows a system to learn from the past or present data set when predictions or decisions are made in the learning process. The implementation of the machine learning tool in a system facilitates and speeds the analytical work. It is also able to apply complex mathematical calculations automatically to solve problems without requiring any machine learning techniques or expertise. Machine learning tool that used in this study is WEKA.

3.4.3.1 WEKA

WEKA stands for Waikato Environment for Knowledge Analysis. It is a popular machine learning software program developed in Java at Waikato University, New Zealand (Thakur, 2015). WEKA supports a number of standard data mining tasks, including data pre-processing, clustering, classification, regression, visualization and selection of features (Science, 2016). This is free software licensed under the (GNU) General Public License. The data file usually used by WEKA is in ARFF file format, which consists of special tags indicating different things in the data file, such as attribute names, attribute types, attribute values and the data (Purva Sewaiwar, 2015). It consists of visualization tools comprising different types of algorithms, such as random forests, multi-layer perceptrons, k-nearest neighbors and regression. Figure 9 illustrates the graphic user interface (GUI) of the WEKA.

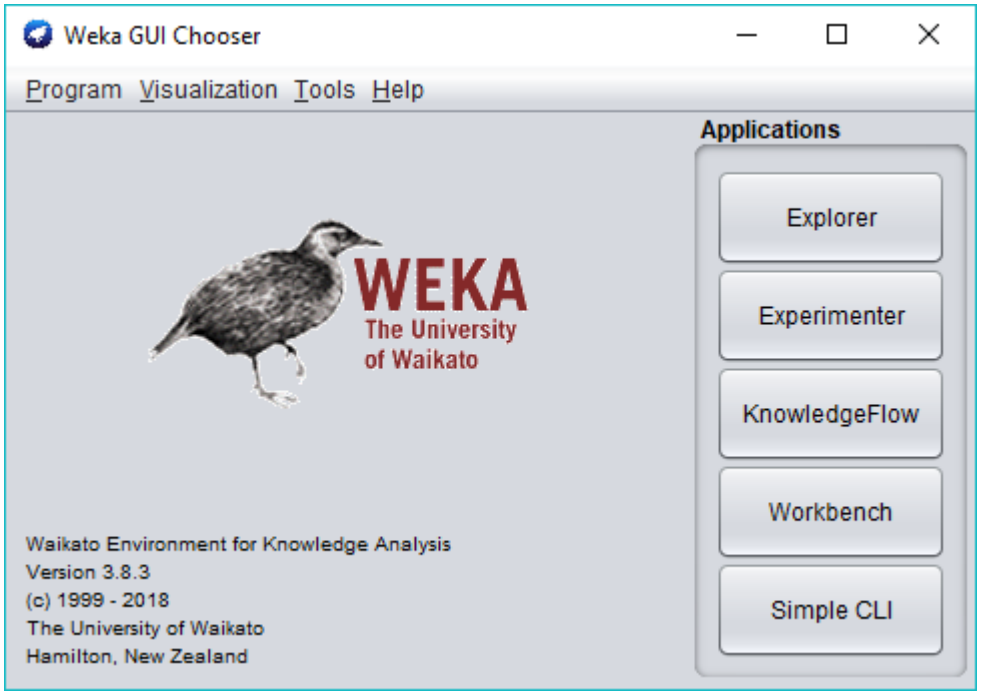


Figure 7 GUI of WEKA

GUI of WEKA has four buttons under Applications:

a. Explorer: It is the main WEKA interface. It has a set of panels, each of which can be used to perform a certain task. One of the other panels in the Explorer can be used for further analysis once a data set has been loaded.
b. Experimenter: An environment for experimentation and statistical testing between learning programs.
c. Knowledge Flow: This environment mainly supports the same functionality as the Explorer but with a drag and drop interface. One advantage is that it supports progressive learning.
d. Simple CLI: Provides a simple command - line interface that directly executes WEKA commands for operating systems that do not have a command - line interface of their own (Purva Sewaiwar, 2015)

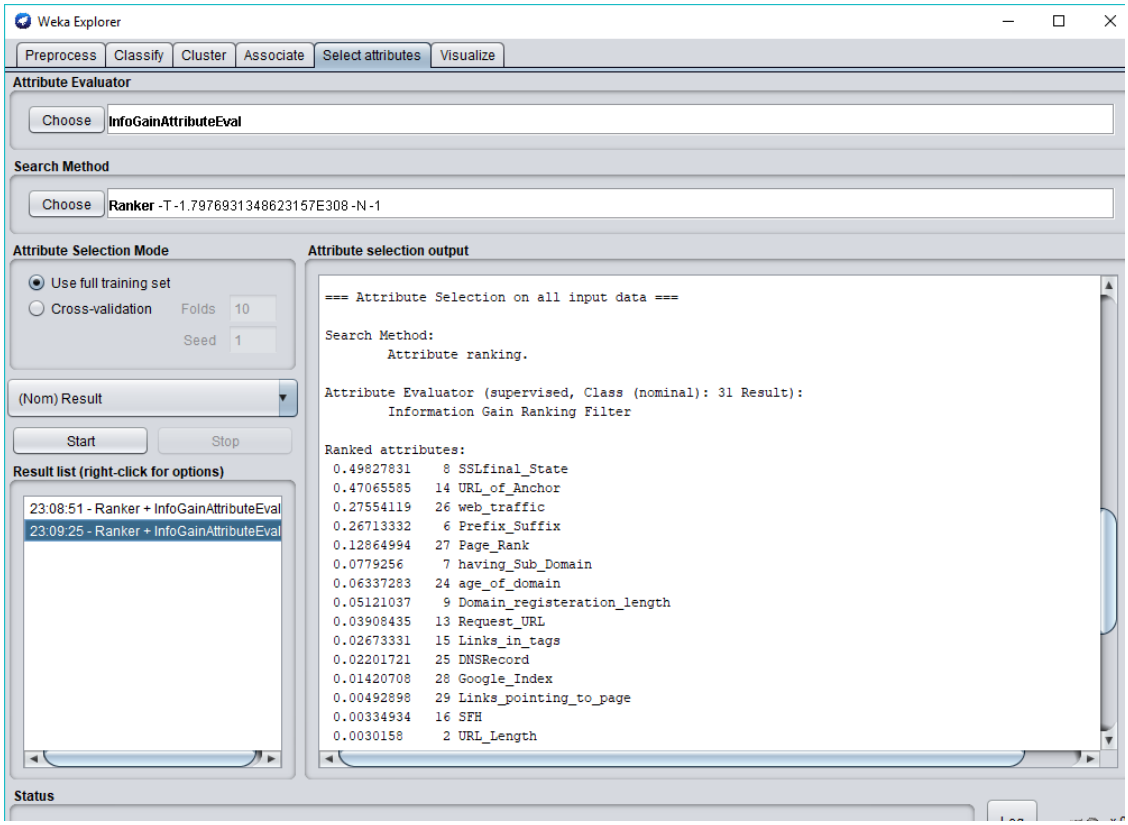


Figure 8 Application for features selection

Figure above shows the information gain from the complete training set with ranker attributes for selecting the relevant features for phishing website detection. With a good interface, even normal users use this application. All algorithms for the selection of classifiers and features are found in this application. The application show good quality results as well as easy to understand.

3.5 Design and Implementation

After the framework has been developed, we need to prove whether the proposed framework acceptable or unacceptable. Therefore, before we implement the system, a procedure was designed to test the accuracy of the anomaly detection method. Before we proceed with the phishing website detection (PWD), the design procedure as shown in figure 3.4 was developed to test theory.

The model of design consists of five components that are collect data, define phishing features, create model, testing and finally the result will be compared. Each component is briefly discussed in the next sub-topic.

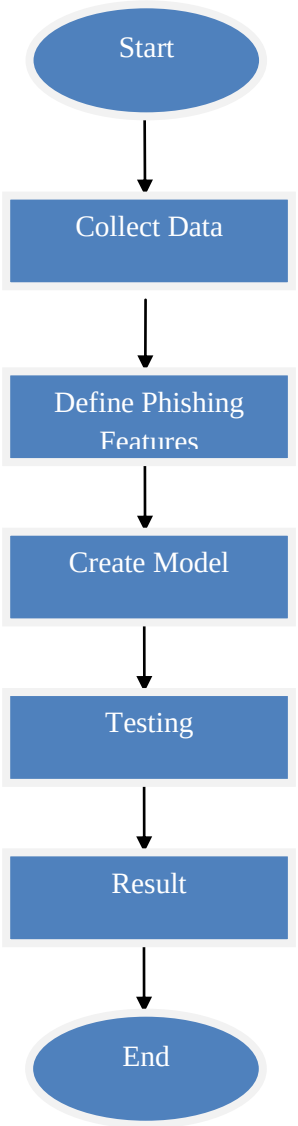


Figure 9 Procedures for Improving Detection Method

The final phase of the project is the implementation phase. In this phase, the design model is used as guideline for implementing the proposed solution. The first step in this phase is to prepare a laptop or a personal computer and software such as WEKA for the project.

The data set is then requested from the internet or from individuals who are willing to share their data for completing the project. The project then proceeds in accordance with the flow designed during the design phase.

3.6 Hardware and Software

In order to complete the overall project, it is necessary to list down all of the requirements needed during project development. In order to carry out the research experiment we must specify the hardware and Software requirements to be used in setting up the experiment. This step is crucial for research, since software and hardware are used to conduct the research experiment and to test and evaluate the experiment in the next phase.

3.6.1 Hardware Requirement

Table 3.1 Hardware Requirement and Purpose

HARDWARE	PURPOSE
1) One unit of Laptop -Processor: Intel(R) Core(TM) i3-3217U CPU @ 1.80GHz, 1801 Mhz, 2 Core(s), 4 Logical Processor(s) -RAM: 2.00GB -System type: x64-based PC	Used for the whole research project, which conducts the research, implementation, testing and documentation of resources.

3.6.2 Software Requirement

Table 3.2 Software Requirement and Purposes

SOFTWARE	PURPOSE
1) Windows 10	- The operating system used in this study
2) Microsoft Excel	- To store the dataset/database
3) Microsoft Word 2016	- For documentation of this project
4) WEKA	- To analyze and optimise the dataset
5) Project Plan 365	- To design a Gantt chart
6) Google Chrome	- To collect information

3.7 Testing and Evaluation

This test and evaluation phase will be the final step in carrying out this study. The experiment will be tested for this stage, as all components are combined. Testing and evaluation is conducted to solve the problem statement and to determine whether the limitation of existing journals is managed to avoid. The main purpose of this test is to demonstrate the proposed the best detection model in order to ensure the accuracy of the results and claims made in this investigation. In addition, the testing and evaluation phase allows the research experiment to identify errors and limitations so that further improvements can be made to obtain the desired result.

Finally, the study, which clearly describes the entire process of this research, is completed. The results are also discussed and recorded to demonstrate whether or not

the objectives are being achieved. In the next chapter we will find a more detailed explanation for the implementation phase.

CHAPTER 4

IMPLEMENTATION, RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, it will be the implementation of the methodology, planning, analysis, and design that has been arranged and draft in Chapter 3. The implementation stage is very crucial during the whole process of developing the tools. It is because this stage will confer on the process of detecting phishing websites using the tools.

4.2 Dataset Description

The first part of implementation is to collect dataset. The dataset phase is important for maintaining result accuracy. The dataset will gives more understanding and explanation of phishing and legit activities. For further examination, the dataset is then analysed and the results are used to foresee or predict the future events in phishing.

All the features were collected from (Mohammad, McCluskey, & Thabtah, 2012). There are a total of 30 phishing website features that has been collected. This dataset collected mainly from a well-known phishing database, PhishTank archive, MillerSmiles archive and Google searching operators. The collected dataset holds categorical values those are “Legitimate”, ”Suspicious” and “Phishy”, these values have been transformed to numerical values by replacing the values “1”, “0” and “-1” instead of “Legitimate”, “Suspicious” and “Phishy” respectively.

4.3 Machine Learning Approach

Machine learning approach is used to ensure that website users are able to optimise the phishing features through the feature optimisation approach. This approach provides shorter training and testing time thus it simplify the phishing detection system.

Feature selection methods were used to identify and remove irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model (T. Chou & Pickard, 2018) .

The features of the phishing website were first trained and then classified by using significant features. In order to choose the significant features for effective phishing website detection, this study applies the feature selection approach. Hence, the number of phishing features was reduced from 30 features to 15 features only. This is to ensure that there is a unique pattern appearing between the normal and phishy websites. Table 4.1 presents the list of phishing website features used by the study.

Table 4.1 Phishing Website Features

Phishing Features	Description
SSLFinal_State	SSL Certificates are small data files that digitally bind a cryptographic key to an organization's details. When installed on a web server, it activates the padlock and the https protocol and allows secure connections from a web server to a browser.
URL_of_Anchor	An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL".
Website Traffic	This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit.
Prefix_Suffix	The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they

	are dealing with a legitimate webpage.
Page_Rank	PageRank is a value ranging from “0” to “1”. PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage.
Having_Sub_domain	A subdomain is a domain that is a part of a larger domain under the Domain Name System (DNS) hierarchy. It is used as an easy way to create a more memorable Web address for specific or unique content with a website.
Age_of_domain	This feature can be extracted from WHOIS database (Whois 2005). Most phishing websites live for a short period of time.
Domain_registration_length	Based on the fact that a phishing website lives for a short period of time, it was believed that trustworthy domains are regularly paid for several years in advance.
Request_URL	Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.
Links_in_tags	It is common for legitimate websites to use <Meta> tags to offer

	metadata about the HTML document; <Script> tags to create a client side script; and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage.
DNSRecord	DNS records are basically mapping files that tell the DNS server which IP address each domain is associated with, and how to handle requests sent to each domain
Google_Index	This feature examines whether a website is in Google's index or not.
Links_pointing_to_page	The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain (Dean, 2014).
SFH	Server Form Handler (SFH) that contain an empty string or "about:blank" are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.
URL_Length	To ensure accuracy of our study, it has been calculated the length of URLs in the dataset and produced an average URL

	length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

4.4 Evaluation and results

The initial results shows the outcomes which were obtained from four machine learning classifiers that are random forest, J48, MLP and KNN. This study also used the parameters of accuracy, FPR, precision, recall and f-measure to investigate the different measurements. Table 4.2 shows the results achieved from 15 phishing website features of testing set which used four selected classifiers.

Table 4.2 Performance of each classifiers

Classifiers	Accuracy (%)	FPR	Precision	Recall	F-measure
Random Forest	94.79	5.3	94.8	94.8	94.8
J48	93.93	6.0	94.0	93.9	93.9
MLP	93.28	7.0	93.3	93.3	93.3
KNN	93.08	6.8	93.1	93.1	93.1

The results indicated that random forest classifiers had achieved a highest accuracy result of 94.79 percent when compared to KNN which achieved only 93.08 percent. This outcome shows that the random forest classifiers are more effective than other selected classifiers in detecting phishing website. It also shows that feature selection plays a crucial role in determining the effectiveness of phishing website detection. The high precision rate shows that the classifier produced more relevant results and producing accurate results.

4.4.1 Confusion matrix

A confusion matrix is a technique for summarizing the performance of a classification model. The table shows two possible classes' prediction, normal and phishing. For example, if a model predicts the presence of phishing activities, the result

will show “phishing” and vice versa. Table 4.3 shows the performance of the four classifiers.

Table 4.3 Confusion matrix of classifiers

Classifiers	Actual	Predicted	
		Predicted normal	Predicted phishing
Random Forest	Actual normal	1032	62
	Actual phishing	66	1296
J48	Actual normal	1033	61
	Actual phishing	88	1274
MLP	Actual normal	1005	89
	Actual phishing	76	1286
KNN	Actual normal	1023	71
	Actual phishing	99	1263

The table above shows that the study produced corrected and magnificent results by predicting the unknown phishing with 1033 for the J48 classifiers. In the incorrectly predicted perspective, the J48 shows the most minimal value. Hence, the outcomes shows that J48 classifiers able to predict unknown phishing more accurately.

4.4.2 Receiver operating characteristics curve (ROC)

In this study, based on the phishing website features, the processes were classified as normal and phishing. Aside from using performance matrix, this study also calculated the receiver operating characteristics (ROC) curve for each of the machine learning classifiers. In this phase, the TPR was regarded as the detection rate which will correctly predicted the phishing process and the FPR was selected as the detection rate

which incorrectly predicted normal as phishing. Figure 4.1 presents the curve for machine learning classifiers.

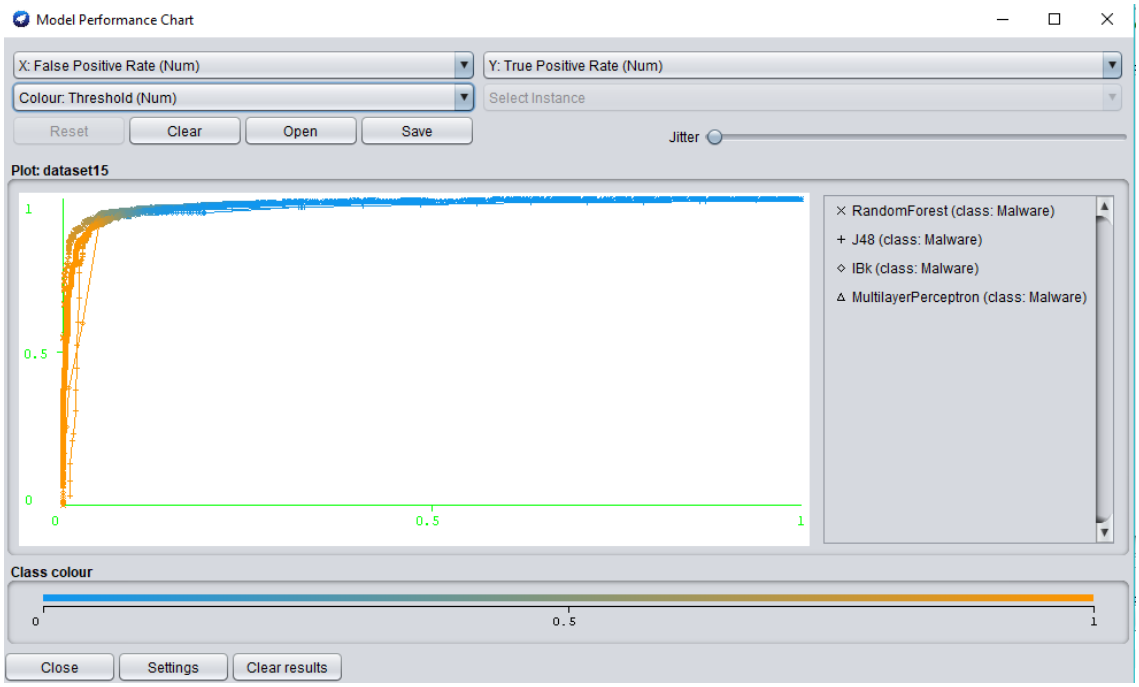


Figure 10ROC Curve

The horizontal axis in the above figure indicates the error detection rate meanwhile the vertical axis indicate the detection rate. Four lines represent the individual ROC curve of the machine learning classifiers. The ROC curve is difficult to compare because it seems to be similar under the same conditions. Hence, the area under the curve (AUC) was used to measure detection accuracy. The AUC results identified were able to measure whether the detection approach was good or bad. An area of 1 shows perfect prediction while an area of 0.5 shows a bad prediction. Table 4.4 shows the AUC performance.

Table 4.4 AUC results

Classifier	AUC	Indicator
Random Forest	0.985	Perfect prediction
KNN	0.961	Perfect prediction

MLP	0.978	Perfect prediction
J48	0.957	Perfect prediction

Table 4.4 shows that the random forest and MLP classifiers provide the best AUC values, with over 0.97. This signifies perfect prediction. Overall, the ROC and the AUC values confirmed that the most recent phishing experiments had provided compelling accurate results in the phishing website applications detection.

4.4.3 Threshold

The optimal threshold is the value that best separates the two detections that are related to the phishing and normal features. The threshold value is used to investigate whether the presence of behaviour pattern indicator is normal (0) or phishing (1). The threshold value for random forest, MLP, KNN and J48 are given in table. As the threshold values were obtained based on the real behaviour patterns of the normal and phishing applications, it can be said that the approach used in this study was able to detect phishing with more than 90 percent accuracy rate.

Table 4.5 Optimal threshold

Classifier	Accuracy	Threshold
Random Forest	0.947	0.315
KNN	0.939	0.250
MLP	0.932	0.359
J48	0.931	0.286

Table 4.5 shows the MLP has an optimal threshold of 0.359 with an accuracy of 0.932. This is the point where the phishing is finally detected. In other words, a threshold value of between 0 to 1 needs to be seen in the system in order for the phishing behaviours to be identified.

4.4.4 Robustness

Apart from evaluating effectiveness of the approach, the robustness of the approach for producing more dependable results were also tested. Robustness is the property that characterizes how effective your algorithm is while being tested on the new independent (but similar) dataset. In the other words, the robust algorithm is the one, the testing error of which is close to the training error. Table 4.6 shows the result of the classifiers' performance.

Table 4.6 Performance Result

Classifiers	Accuracy (%)	FPR	Precision	Recall	F-measure	ROC
Random Forest	94.79	5.3	94.8	94.8	94.8	98.5
J48	93.93	6.0	94.0	93.9	93.9	95.7
MLP	93.28	7.0	93.3	93.3	93.3	97.8
KNN	93.08	6.8	93.1	93.1	93.1	96.1

The table shows that the approach applied in this study was able to detect unknown phishing with over 95 percent accuracy rate.

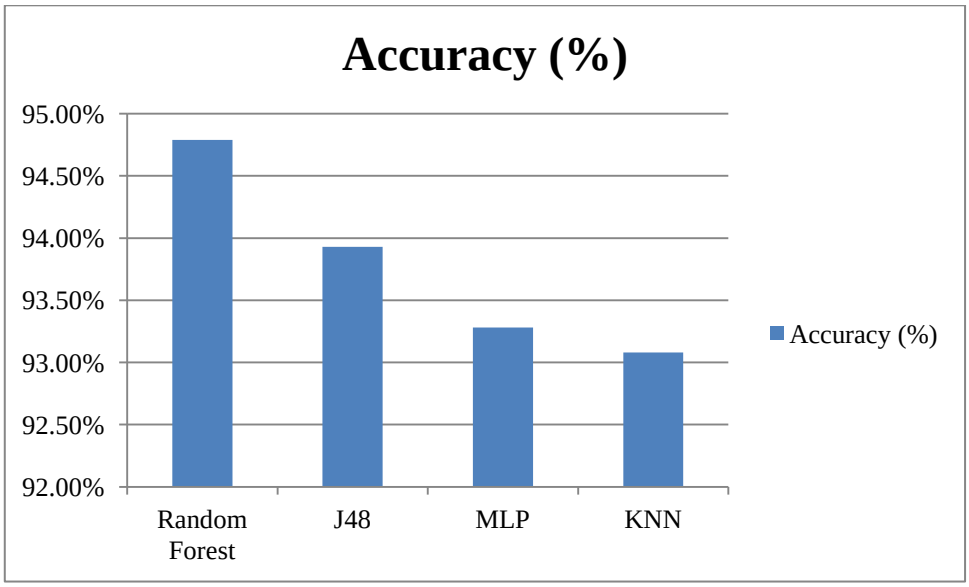


Figure 11 Percentage accuracy

This figure above shows percentage of accuracy of the detection based on the four classifiers. Random forest classifier shows the highest percentage of the accuracy by 94.79% compared to other classifiers. Second highest classifier is J48 by 93.93% and followed by MLP with 93.28% and the last is KNN with 93.08%.

Table 4.7 The accuracy results comparison with past research papers

Classifiers	Accuracy Results	Source
Random Forest	81.80%	(Hodžić & Kevrić, 2016)
	94.79%	This research
KNN	96.18%	(Lee & Kim, 2015)
MLP	89.00%	(Hodžić & Kevrić, 2016)
	93.28%	This research
J48	73.90%	(Hodžić & Kevrić,

		2016)
	93.93%	This research

Table 4.7 shows the comparison of the accuracy results for the algorithms tested in this research with the previous research papers. The results recorded are the highest results compared to other algorithms. Based on the table, the highest accuracy result for the paper was for KNN algorithm, with 96.18% compare to other paper as well as this paper. Then, for Random Forest, MLP and J48 this paper recorded as the highest accuracy results in this research compared to the previous research papers.

Table 4.8 Time taken to produce model (seconds)

Classifiers	Build model
Random Forest	1.61
J48	0.27
MLP	13.47
KNN	0.02

Table 4.8 shows the time taken to produce the results in second. The results show that KNN has the lowest model complexity as it uses minimal time to build the model. Table 4.5 presents the robustness of the approach based on the time taken to produce the model. Hence, it can be concluded that to achieve reasonable accuracy and effectiveness in classifying unknown phishing, robustness is critical as it helps to determine the performance of the classifiers.

CHAPTER 5

CONCLUSION

5.1 Introduction

Today, the internet has changed the way of live for human. There are wide range of activities from searching for information to entertainment, online shopping, financial services and also socializing. Frequent usage of internet makes people have come to trust the Internet to provide gateway for office, home and personal convenience.

Online transactions nowadays are becoming more relevant and provide the easiest and fastest way to managing and handling things. There is nothing impossible to be done quicker and simplest by having Internet. Despite the advantages and benefits provided, it has to be its own disadvantages and that is security. Many people rarely realize about this security issues which may bring harm to them.

This study provides understanding about phishing. This study also aims to detect phishing website by using machine learning. The dataset of phishing features are collected and they have been through feature optimization approach. This approach makes the list of phishing features lesser and provides smaller dataset. Then, it applies machine learning classifiers that are Random Forest, J48, MLP and KNN. The parameters are taken account into in order to detect phishing website effectively.

5.2 Research Objectives

The purpose of this study was to improve a phishing website detection system by using machine learning for website URL. Section 1.3 had described the three research objectives of this study.

Objective 1: To investigate security flaws by analyzing the state-of-the-art phishing detection system

The first objective was to investigate the security vulnerabilities by analysing the current existing study on phishing website detection system. The research objective was achieved through a thorough review of the most important works published in online scholarly journals. This objective was achieved through Chapter 2, in which all information concerning the phishing website detection system was presented. Chapter 2 also presented the classification of phishing website detection and machine learning approach as well as the algorithms.

Objective 2: To propose a phishing detection system that analyzes website applications using machine learning

The second research objective was to evaluate the phishing website detection system based on the machine learning approach. The evaluation of the phishing website detection system was examined by using WEKA. The experiments tested the features in six evaluation measures in the WEKA simulation: accuracy, False Positive Rate (FPR), True Positive Rate (TPR), precision, recall, and f-measure. This objective was achieved in Chapter 4.

Objective 3: To evaluate the proposed system in terms of accuracy of detection.

The third objective is to evaluate the proposed system in terms of detection accuracy. This has been done by evaluating the performance by the four classifiers. Based on the result, random forest shows the highest percentage of accuracy to detect phishing website.

5.3 Achievement of the study

This research began by studying the evolution of phishing and examining the various types of phishing website detection systems. It examined the issues concerning the detection of phishing websites and the selection of relevant features. Several machine learning classifiers have been examined and performance results have been collected. The study evaluated the results in order to fulfil the objective of the study. As noted below, several points of interest have been identified.

5.3.1 A detection model for phishing

This study has developed a model that can detect phishing websites by means of a static analysis. An approach to machine learning has been used as a better adaptive detection model. The model worked very well to detect phishing website based on the given data set.

5.3.2 Issues in phishing website detection studies

In chapter 2, this study presented the phishing website detection types and their relevance in phishing detection. Several strategies to address the limitations have been identified by presenting the strengths and weaknesses of these problems. In order to improve the efficiency of the phishing website detection system, research has been carried out to highlight some of the limitations. The objective was to look for the relevant features that developed a more efficient approach.

5.3.3 Issues in phishing website feature selection

This study has shown a critical analysis of the different perspectives used to address the major problems of selecting features, with the aim of improving detection performance and minimizing complexity.

5.4 Research Constraints

The discussions in the previous chapters have confirmed that this research has satisfactorily achieved its aims and objectives. However, there were a number of constraints and obstacles in the study that are mentioned here for future references.

5.4.1 Sample size

The sample size used is small, so it was difficult to identify significant relationships from the data. The number of analytical samples being used in this study has an impact on this research, as statistical tests usually require a larger sample size to ensure a representative distribution of the population.

5.4.2 The assessment of the study was carried out using a static detection model only

In this study, all input features are collected from static analysis. Nonetheless, in the practical solution both static and dynamic have their own advantages and disadvantages. A comparison of the results of both analyses would therefore be more useful.

5.4.3 Time

The time available to investigate a research problem and measure change or stability over time is quite limited by the due date of the task.

5.5 Future works

The following recommendations for future work outside the scope of this study were listed as follows:

5.5.1 Selection of relevant features

The more complex and extensive data becomes, the harder it becomes to choose relevant and suitable features to improve detection performance. The process requires further analysis to investigate the correlation between malware and benign applications. This will reduces false alarms, thus increase the detection accuracy.

5.5.2 Enhance false alarm rate

False alarm rate remains a problem as long as it exists in the detection module. False alarms refer to the statistical measurement of how well the sample dataset classifies the phishing website correctly. This means that the phishing data was incorrectly predicted as normal. This problem leads to incorrect detection of websites

and even small amounts of false alarms can cause enormous impacts. A reliable and efficient detection module is therefore needed to solve this problem.

5.5.3 Dynamic analysis approach

This study also can be done by using Dynamic Analysis Approach. It can identify vulnerabilities in a runtime environment. This approach recognises vulnerabilities that could have been false negatives in static code analysis.

REFERENCES

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit on - ECrime '07*, 60–69. https://doi.org/10.1145/1299015.1299021

Amalina, F., Ali, N., Badrul, N., & Abdullah, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, 343–357. https://doi.org/10.1007/s00500-014-1511-6

APWG. (2018). *Phishing Activity Trends Report 1 Quarter*. *Most* (Vol. 1).

Chaudhry, J. A., Chaudhry, S. A., & Rittenhouse, R. G. (2016). Phishing attacks and defenses. *International Journal of Security and Its Applications*, *10*(1), 247–256. https://doi.org/10.14257/ijsia.2016.10.1.23

Chou, N., Ledesma, R., Teraguchi, Y., Mitchell, J. C., & Ca, S. (2004). Client-side defense against web-based identity theft. *Ndss*, 1–16. https://doi.org/10.1.1.65.679

Chou, T., & Pickard, J. (2018). Machine Learning based IP Network Traffic Classification using Feature Significance Analysis, *16*(3), 9–12.

Hodžić, A., & Kevrić, J. (2016). Comparison of Machine Learning Techniques. *ICESoS 2016 - Proceedings Book*, 249–256. https://doi.org/10.1109/WETICE.2011.28

Lee, J., & Kim, D. (2015). Heuristic-based Approach for Phishing Site Detection using URL Features.pdf, 131–135. Retrieved from http://eprints.ibu.edu.ba/3308/1/Adnan Hodzic Jasmin Kevric and Adem Karadag.pdf

Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2009). An evaluation of machine learning-based methods for detection of phishing sites. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5506 LNCS*(PART 1), 539–546. https://doi.org/10.1007/978-3-642-02490-0_66

Netcraft Ltd. (2008). Netcraft Toolbar. Retrieved from https://toolbar.netcraft.com/

Nisha, S., & Madheswari, A. N. (2016). SECURED AUTHENTICATION FOR INTERNET VOTING IN CORPORATE COMPANIES TO PREVENT PHISHING ATTACKS, *22*(1), 45–49.

Purva Sewaiwar, K. K. V. (2015). Comparative Study of Various Decision Tree Classification Algorithm Using WEKA, *9359*(10), 87–91.

Science, C. (2016). COMPARATIVE EVALUATION OF THE DIFFERENT DATA MINING TECHNIQUES, *10*(3), 233–238. https://doi.org/10.1515/ama-2016-0036

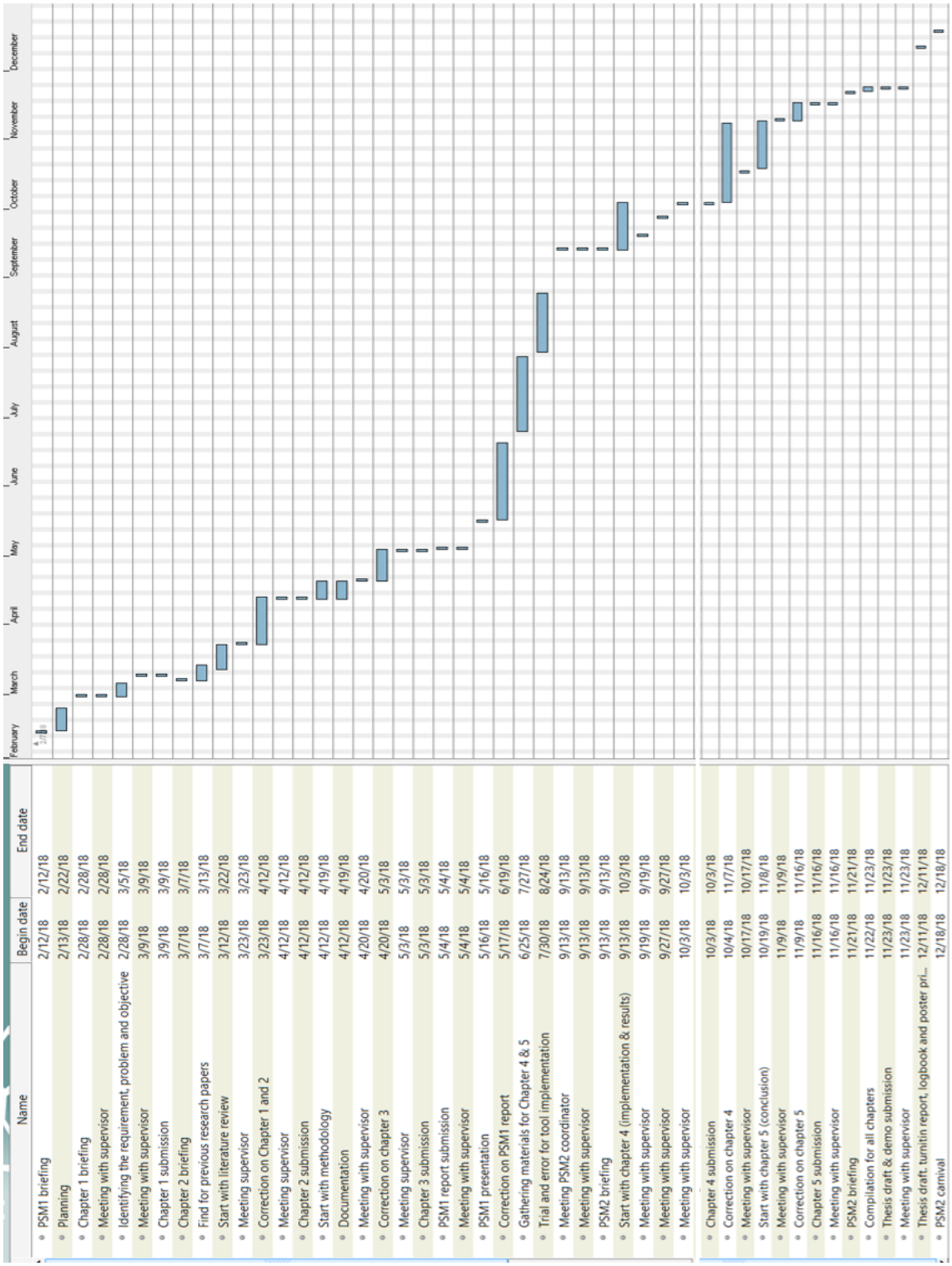
Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C. (2009). An Empirical Analysis of Phishing Blacklists. *6th Conference on Email and Anti-Spam*, (March 2014). Retrieved from http://repository.cmu.edu/hcii%5Cnhttp://repository.cmu.edu/hcii/282

Thakur, R. (2015). Preprocessing and Classification of Data Analysis in Institutional System using Weka, *112*(6), 9–11.

Vanhoenshoven, F., Gonzalo, N., Falcon, R., Vanhoof, K., & Mario, K. (2016). Detecting Malicious URLs using Machine Learning Techniques. https://doi.org/10.1109/SSCI.2016.7850079

Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). Cantina+. *ACM Transactions on Information and System Security*, *14*(2), 1–28. https://doi.org/10.1145/2019599.2019606

APPENDIX A



Name	Begin date	End date
PSM1 briefing	2/12/18	2/12/18
Planning	2/13/18	2/22/18
Chapter 1 briefing	2/28/18	2/28/18
Meeting with supervisor	2/28/18	2/28/18
Identifying the requirement, problem and objective	2/28/18	3/5/18
Meeting with supervisor	3/9/18	3/9/18
Chapter 1 submission	3/9/18	3/9/18
Chapter 2 briefing	3/7/18	3/7/18
Find for previous research papers	3/7/18	3/13/18
Start with literature review	3/12/18	3/22/18
Meeting supervisor	3/23/18	3/23/18
Correction on Chapter 1 and 2	3/23/18	4/12/18
Meeting supervisor	4/12/18	4/12/18
Chapter 2 submission	4/12/18	4/12/18
Start with methodology	4/12/18	4/19/18
Documentation	4/12/18	4/19/18
Meeting with supervisor	4/20/18	4/20/18
Correction on chapter 3	4/20/18	5/3/18
Meeting supervisor	5/3/18	5/3/18
Chapter 3 submission	5/3/18	5/3/18
PSM1 report submission	5/4/18	5/4/18
Meeting with supervisor	5/4/18	5/4/18
PSM1 presentation	5/16/18	5/16/18
Correction on PSM1 report	5/17/18	6/19/18
Gathering materials for Chapter 4 & 5	6/25/18	7/27/18
Trial and error for tool implementation	7/30/18	8/24/18
Meeting PSM2 coordinator	9/13/18	9/13/18
Meeting with supervisor	9/13/18	9/13/18
PSM2 briefing	9/13/18	9/13/18
Start with chapter 4 (implementation & results)	9/13/18	10/3/18
Meeting with supervisor	9/19/18	9/19/18
Meeting with supervisor	9/27/18	9/27/18
Meeting with supervisor	10/3/18	10/3/18
Chapter 4 submission	10/3/18	10/3/18
Correction on chapter 4	10/4/18	11/7/18
Meeting with supervisor	10/17/18	10/17/18
Start with chapter 5 (conclusion)	10/19/18	11/8/18
Meeting with supervisor	11/9/18	11/9/18
Correction on chapter 5	11/9/18	11/16/18
Chapter 5 submission	11/16/18	11/16/18
Meeting with supervisor	11/16/18	11/16/18
PSM2 briefing	11/21/18	11/21/18
Compilation for all chapters	11/22/18	11/23/18
Thesis draft & demo submission	11/23/18	11/23/18
Meeting with supervisor	11/23/18	11/23/18
Thesis draft, turnitin report, logbook and poster pri...	12/11/18	12/11/18
PSM2 carnival	12/18/18	12/18/18

Gantt Chart