PHISHING ATTACK DETECTION USING MACHINE LEARNING METHOD

JOHN ARTHUR JUPIN

Bachelor of Computer Science
(Computer Systems and Networking) with Honors

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : JOHN ARTHUR JUPIN

Date of Birth :

Title : PHISHING ATTACK DETECTION USING MACHINE LEARNING METHOD

Academic Session : 2018/2019

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)

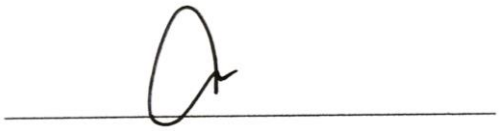
New IC/Passport Number
Date: 8 JANUARY 2019

DR. MOHD ARFIAN BIN ISMAIL
Name of Supervisor
Date: 8 JANUARY 2019

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

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 Bachelor of Computer Science (Computer System and Networking).



(Supervisor's Signature)

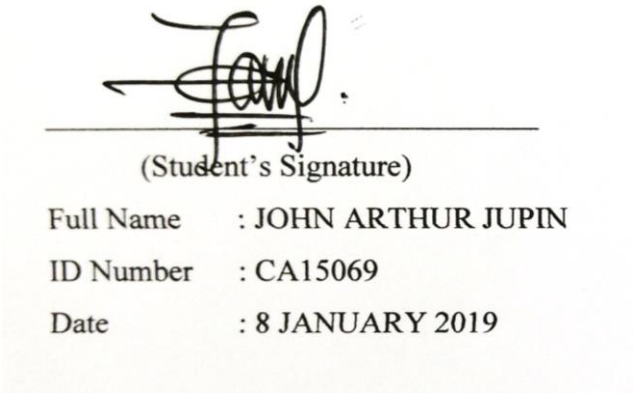
Full Name : DR. MOHD ARFIAN BIN ISMAIL

Position : SENIOR LECTURER

Date : 8 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 : JOHN ARTHUR JUPIN

ID Number : CA15069

Date : 8 JANUARY 2019

PHISHING ATTACK DETECTION USING MACHINE LEARNING METHOD

JOHN ARTHUR JUPIN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science (Computer Systems and Networking) with Honors

Faculty of Computer Science & Software Engineering

UNIVERSITI MALAYSIA PAHANG

JANUARY 2019

ACKNOWLEDGEMENTS

First, I would like to thank to my supervisor, Dr. Mohd Arfian bin Ismail, Senior Lecturer of Faculty of Computer Science and Software Engineering of University Malaysia Pahang for helping me from the start of this research until the writing of the thesis for this research. He had given me his full authority on my research yet monitor my research progress. He also had given his advices on how this research should be conducted.

I also would like to thank Prof. Dr. Kamal Zuhairi Zamli as the dean of Faculty of Computer Science and Software Engineering of University Malaysia Pahang for giving students the chance to use the lab during non-academic session.

Finally, I would like to thank my family members, friends and my lectures for supporting and guiding me from the start of this research until the end of this research.

ABSTRAK

Pembangunan rangkaian komputer pada masa kini adalah sangat pesat. Perkara ini tidak dapat lagi dinafikan kerana setiap pengguna komputer di serata dunia perlu menyambungkan komputer mereka ke rangkaian Internet. Hal ini demikian menunjukkan bahawa penggunaan rangkaian Internet adalah sangat penting, sama ada ianya digunakan untuk tujuan kerja dan tugasan mahupun untuk mengakses ke akaun media sosial, contohnya Instagram, Facebook dan Twitter. Walaubagaimanapun, dalam pengunaan luas rangkaian komputer ini, secara tidak langsung, privasi pengguna komputer adalah dalam bahaya. Hal ini demikian disebabkan pengguna komputer tidak menitikberatkan sistem sekuriti dan keselamatan di dalam komputer mereka. Oleh sebab ini, penggodam akan menggodam dan membuat serangan rangkaian ke atas pengguna komputer dengan mudah. Hal ini demikian sangat bahaya, terutama sekali kepada organisasi penting kerana penggodam dapat melumpuhkan sistem atas talian dalam syarikat, mencuri maklumat-maklumat sulit dan seterusnya mencuri wang syarikat secara atas talian tanpa disedari oleh mana-mana pihak. Antara serangan yang boleh dibuat termasuklah serangan penafian-perkhidmatan, serangan perdayaan dan *phishing*. Matlamat dalam kajian ini ialah untuk menggunakan alat *anti-phishing* dalam menghalang serangan sekuriti rangkaian di dalam satu organisasi. Dalam kajian ini, serangan *phishing* telah dikaji secara mendalam. Selepas kajian dibuat, cara yang telah diutarakan untuk menghalang serangan *phishing* ialah melalui pembelajaran mesin. Selain itu, kajian ini juga menunjukkan bahawa serangan *phishing* selalunya berhubung kait dengan serangan mesej *spam*. Mesej *spam* ini termasuklah email dan juga mesej SMS yang diterima dari pengguna. Dalam pembelajaran mesin, terdapat beberapa algorithma yang boleh digunakan dalam menghalang kedua-dua serangan ini. Algorithma *Naïve Bayes*, algorithma Pokok Keputusan dan algorithma Mesin Vektor Sokongan telah digunakan untuk menghalang serangan *spam*, dan juga serangan *phishing* daripada berlaku. Kajian algorithma ini dibuat secara mendalam dan cara-cara dalam melaksanakan algorithma ini telah dibincangkan secara mendalam dan lebih terpencil. Eksperiment juga telah dijalankan untuk set data yang diperoleh dengan menggunakan kaedah pembelajaran mesin. Keputusan telah diperoleh, di mana menunjukkan prestasi untuk kaedah pembelajaran mesin untuk setiap set data.

ABSTRACT

The development of computer networks today is increased rapidly. This can be shown based on the trend of every computer user around the world, whereby they need to connect their computer to the Internet. This shows that the use of Internet networks is very important, whether they used it for work and assignment purposes, or for the access to social media accounts, such as Instagram, Facebook and Twitter. However, in this wide use of this computer network, the privacy of computer users is in danger. This is because some of the computer users do not install security system in their computer. This problem will allow the hackers to hack and commit the network attacks. This is very dangerous, especially to the important organizations because hackers can disable the online system in the company, steal confidential information and subsequently steal company money through online without being aware of any one. The attacks that can be made includes denial-of-Service attack, DNS spoofing attack and phishing attack. The goal of this study is to apply anti-phishing tools in preventing the network security attack in an organization. In this study, phishing attacks have been studied thoroughly. After a study has been made, machine learning method is used to prevent the phishing attack. Besides, the study also shows that phishing attack is always related to the spam attack, where there might be attached phishing link in the spam message. This spam message includes the email and the SMS message that received by the user. There are several algorithms that can be used in the machine learning method to prevent the both attacks. The Naïve Bayes algorithm, Decision Tree algorithm and Support Vector Machine algorithm has been used to prevent the spam attack, as well as the phishing attack. The study of this algorithm is made thoroughly and the methods in implementing this algorithm have been discussed in detail. The experiment is conducted for the datasets that obtained by using machine learning method. The results are obtained, showing the performance of machine learning method on each dataset.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNS	Domain Name System
URL	Uniform Resource Locator
SVM	Support Vector Machine
ARFF	Attribute Relation File Format

CHAPTER 1

INTRODUCTION

1.1 Background of study

Network security is the most important and critical issues that need to be considered and emphasized in the network, especially in an organization, such as offices, banks and clinics. Basically, network security is the authorization, commonly by using a username and password, which inhibit and monitor the unauthorized access and all the administrator event in the network (Pawar & Anuradha, 2015).

It is important for the organization to maintain their security network to ensure the privacy and confidentiality of their employer data, as well as their employee data. This will ensure the data, especially the sensitive data, such as the employee information details, can be stored in the server safely. For example, for us to access the online banking, we need to have an authentication to access our account. This can be done by providing the username and password in the login page of the online banking. Authentication is needed in this scenario so that our sensitive data would not be exposed to the unauthorized user or the hacker.

Although there is implementation of the network security in an organization, but still there is network attack happened. The network attack that usually happen includes phishing, denial-of-Service attack and Domain Name System (DNS) spoofing. This attack will contribute to the financially and privacy loss to the victims. For example, when the hacker attacks sensitive information while the user using their online banking account, the attacker will use this information to retrieve back the victim's account and then steal their money inside the account. This also can be applied to the office organization, whereby the hacker will gain the sensitive data and use it to commit online crimes, such as stealing the office's money and the data of their employer over the Internet.

Phishing is one of the network security attack, which is the derivational of word 'Fishing' by replacing the 'F' with 'Ph'. Phishing is the act of imitate the genuine websites to collect the sensitive information from the victim and use it for committing crimes, such as illegal financial gain (Kaur & Kaur, 2015). This attack typically starts when the hacker sends an email that seems original to the victim and persuade them to update and verify their information by clicking the Uniform Resource Locator (URL) link in the email (Mohammad, Thabtah, & McCluskey, 2015). Usually, the phishing email will redirect the user to the infected website and asking them to provide their particular information, such as their personal details and bank account information, which will be used to hack the information whatever the user enter (Suganya, 2016). The phishing attack is always related to the spamming email that received by the victim. Those spam emails are also vulnerable to the phishing attack because some of the spam email may contain the link that will redirect the victim to the phishing websites.

The phishing attack can be prevented using the machine learning method. According to Marsland (2015), machine learning is the modification or adaptation of the computer actions so that we can get the more accurate actions in the end. Besides, machine learning is also considered as computational complex since it will lead to produce algorithm. Based on the machine learning method, the prevention of phishing attack can be classified to several algorithm, which includes Decision Tree algorithm, Naïve Bayes algorithm and Support Vector Machine (SVM) algorithm (Smadi, Aslam, Zhang, Alasem, & Hossain, 2016). All the algorithms stated are used in classifying the spam email and the SMS messages datasets. This will show the performance of each of the algorithm, in terms of their accuracy.

1.2 Problem statement

After doing some research, there are some methodologies of overcoming the attack that has been found, which is already exists. Each of them has their own advantages and disadvantages respectively. The summary of the problems is tabulated in the Table 1.1;

Table 1.1: Tabulation of the problem

No.	Problem	Description	Effect
1	Classification of the phishing result is not accurate (Smadi et al., 2016).	The result that obtained after the test, which is the true positive and the false positive does not be considered.	The result that obtained might not be correct and accurate.
2	The level of performance of the method (Smadi et al., 2016).	The performance level of the method does not be considered when the method was used.	The result of the test might take longer time to be obtain.

1.3 Research goal and objective

The goal of this research is to detect phishing attack. The objective of this study is stated as below;

i. To study the issues of phishing attack.
ii. To use machine learning method, which is Naïve Bayes algorithm, Decision Tree algorithm and Support Vector Machine algorithm in detection of phishing attack.
iii. To evaluate the performance of machine learning method in detection of phishing attack.

1.4 Scope of the research

The scope of this research is listed as follow;

i. The research is focus on the method on how to overcome the network security attack, which is machine learning method.
ii. The network security attack will be observed thoroughly.
iii. The dataset that will be use in this algorithm is the email message dataset and SMS message dataset, which contains spam and legitimate (ham) messages.

1.5 Significance of the research

The significance of this research is listed as follow;

i. Can reduce the number of phishing attack towards the computer user.
ii. Give the knowledge and information about the methods that can prevent network security attack, which is phishing attack.
iii. Increasing the awareness of the computer user with the network security issues, especially in the organization.

1.6 Report organisation

The structure of the research can be outlined as follow;

i. Chapter 1: This chapter is a phase where before the other parts is explained, the general information about this research is stated. Background, problem statement, goal, objectives, scope and significance of this research is also being discussed. Basically, this chapter is important as the root from all chapter is based on the content inside it.
ii. Chapter 2: In this chapter, the discussion on existing methods were discussed as well as the comparison of advantages and disadvantages between them. The details on the network security attack are also analysed and studied.
iii. Chapter 3: Discussion in this chapter start with the research methodology which covered four fundamental steps followed by the review of datasets involved. The experimental environment including hardware and software that were used also studied.
iv. Chapter 4: This chapter will discuss on the results of the experiment that tested on the datasets. Besides, the discussion is made thoroughly on the results of finding based on the experiment that have been done.
v. Chapter 5: This chapter concludes the research. The discussion on the constraints and future work also was presented in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will discuss on the existing method that has been widely used to overcome the network attack in organization. The type of attack that will be focused in this study is phishing attack. Besides, machine learning method will be focused as the proposed method in this study. Based on machine learning method, there are three existing algorithms that have widely been used to overcome phishing attack, which includes Naïve Bayes algorithm, Decision Tree algorithm and Support Vector Machine (SVM) algorithm. A literature review in this study was presented from the basic information of all algorithms including the advantages and disadvantages of each methods.

2.2 Phishing attack

Phishing is one of the network security attack that is happen nowadays. It is the process where the attacker, which is the phisher, trying to get the sensitive information from the victim, by pretending that they are from the trusted organization in the Internet. According to Rathod & Pattewar (2015), phishing attack usually will be dispersed through the spam email. This email will be used by the phisher to steal the money and damage the social reputation of the victims. However, there are many methods that can be used to prevent the phishing attack. In this case, the method that will be choose is machine learning method. The are several types of machine learning method that will be discussed in this chapter, which includes Naïve Bayes algorithm, decision tree algorithm and Support Vector Machine (SVM) algorithm.

2.3 Existing Machine Learning Method

There are three types of classifications of machine learning method that will be discuss. The three types of existing method which stated above are explain thoroughly as below.

2.3.1 Naïve Bayes algorithm

Naïve Bayes algorithm, which also known as Bayesian classifier is a group of classification algorithm based on the Bayes Theorem. This classification will shares common principle, where every characteristic that being classified is independent of its value among any other characteristics (Waldron, 2015). This algorithm will calculates a set of probability based on the combination and frequency of the values (T. R. Patil, 2013). According to Kumar & Chaudhary (2017), the general equations for Bayes theorem can be expressed in equation (1);

$$P(x|Y)=\frac{P(Y|x)P(x)}{P(Y)} \tag{1}$$

where;

P(x) : independent probability of x: prior probability

P(Y) : independent probability of Y

P(Y|x) : conditional probability of Y given h: likelihood

P(x|Y) : conditional probability of x given Y

According to Rathod & Pattewar (2015), Naïve Bayes classifier will use text classification method to filter the spam email of the victim. It will use the tokens, which represents as the words that use in the spam and non-spam email to calculate the probability of the email whether it is spamming email or not. Thus, the concept for this algorithm can be expressed in the equation (2), (3), (4), (5), (6) and (7);

$$\text{Prior probability of legitimate email} = \frac{\text{Number of legitimate email}}{\text{Total number of email}} \tag{2}$$

$$\text{Prior probability of spam email} = \frac{\text{Number of spam email}}{\text{Total number of email}} \tag{3}$$

$$\text{Likelihood of } X\text{-email given legitimate} = \frac{\text{Number of legitimate email in the vicinity of } X\text{-email}}{\text{Total number of legitimate email}} \quad (4)$$

$$\text{Likelihood of } X\text{-email given spam} = \frac{\text{Number of spam mail in the vicinity of } X\text{-email}}{\text{Total number of legitimate email}} \quad (5)$$

$$\text{Posterior probability of } X\text{ - email being legitimate} = \text{Prior probability of legitimate email} \times \text{Likelihood of } X\text{ - email given legitimate} \quad (6)$$

$$\text{Posterior probability of } X\text{ - email being spam} = \text{Prior probability of spam email} \times \text{Likelihood of } X\text{ - email given spam} \quad (7)$$

Based on the equations above, the classification of the *X*-email whether it is spam or not can be made based on the value of the posterior probability that is obtained. The higher the value of the posterior probability, the more vulnerable the email is, which shows the probability of the email is spamming email.

The performance results that obtained by using this algorithm are measured in terms of accuracy, error, time taken, precision and recall. There are also three datasets that used which is 1000 mails, 1500 mails and 2100 mails. Table 2.1 shows the performance measurement results of three datasets based on the four terms that stated above (Rathod & Pattewar, 2015).

Table 2.1: The performance measurement

Bayesian Classifier	**Accuracy (TP) (%)**	**Error (TN) (%)**	**Time (MS)**	**Precision**	**Recall**
Dataset 1 (1000 mails)	93.98	6.02	7834.0	0.93	0.95
Dataset 2 (1500 mails)	94.85	5.15	12294.0	0.93	0.81
Dataset 3 (2100 mails)	96.46	3.54	16546.0	0.95	0.87

TP : True Positive

TN : True Negative

Based on the performance measurement results, we can conclude that the higher the number of datasets, the higher the precision and accuracy percentage, the smaller the percentage of error rate. However, the time taken of the experiment to be completed will be longer as the number of datasets increases.

2.3.2 Decision tree algorithm

Decision tree algorithm is the algorithm that belongs to supervised classification algorithm. This algorithm is used in solving the regression and classification problems and used to create a training model, which will predict class or value of target variables that summarized from the training data (Saxena, 2017a).

Decision tree can be implemented by using some types of the algorithms. This includes Iterative Dichotomiser 3 (ID3) algorithm and C4.5 algorithm. According to Kozak & Boryczka (2016), ID3 will utilize the process for creating a decision tree in the "top-down" form. It has been proven a very useful method, but still it has huge number of constraint, which will cause this algorithm is inapplicabe in many real world situations. The C4.5 algorithm was develop to overcome this problem, and has been considered as the good solution when using a large size, missing and continuous variables data.

C4.5 algorithm is a sucessor of ID3 algorithm and a decision tree algorithm that used to detect phishing websites, which are usually found attached inside the spam email. This algorithm are categorized as classification algorithm, which will involves two steps. The two steps that involved includes learning step and classification step (Akansha & Meenakshi, 2017). According to Yang, Yan, Yang, & Li (2017), the algorithm can be expressed in the equation (8), (9), (10), (11) and (12);

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 \left(p_i \right) \tag{8}$$

$$Info_A \left(D \right) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info\left(D_j \right) \tag{9}$$

$$SplitInfo_A \left(D \right) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right) \tag{10}$$

$$Gain\left(A \right) = Info\left(D \right) - Info_A \left(D \right) \tag{11}$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \tag{12}$$

where;

D : training set of class-labelled tuple

D_j : subset of D

C_i : the class label of tupple (for i = 1, …, m)

p_i : probability that a tuple in D belongs to class C_i

|D| : the number of tuples in D

According to Akansha & Meenakshi (2017), this algorithm is tested for the phishing detection by using WEKA tools. This test is based on the J48 optimized implementation of C4.5, which will generate a decision tree once the test is completed. The testing dataset that used contains 300 websites. Based on the test, it is found that 200 websites are detected as phishing websites. The success rate and error rate that obtained is 0.826 and 0.173 respectively after prediction confusion matrix is generated. Thus, the accuracy of the classifier model that trained with 750 instances is 82.6%.

This algorithm will provide a better result if there is higher number of rules, which will let the test dataset to be checked more accurately. Based on this statement, we can conclude that the higher the number of instances in training dataset, the more accurate decision tree is generated.

2.3.3 Support Vector Machine algorithm

The SVM algorithm, which is also known as SVM classifier is a machine algorithm which is mostly used in classification problems (Ray, 2017). It is also a supervised learning technique, whereby it will classifies the dataset that contains class labels and features (Saxena, 2017b). According to P. Patil, Rane, & Bhalekar (2017), SVM algorithm is a linear strong classifier, which can identify two classes label in the dataset. This algorithm will produce a set of hyperplanes, which the maximum marginal hyperplane will be considered at the end of the test. The SVM algorithm can be expressed in the equation (13) and (14);

$$\min \frac{1}{2} |w|^2 + c \sum_{i=1}^{n} \xi_i \tag{13}$$

$$y_i \left(w x_i - b \right) >> 1 - \xi_i \qquad \xi_i \geq 0 \tag{14}$$

where;

i = 1, 2, …, n

n : the dimensionality of the feature

x : input vector

w : the normal vector to the hyperplane

C : capacity constant

ξ_i : parameters for handling no separable data (inputs)

Since the phishing websites is usually attached to the spam email, this algorithm will be suitable to help for the detection of the phishing websites that might attached in the mail. There are some attributes that used in this algorithm to detect the phishing websites. Table 2.2 shows the tabulation of the attributes and the significance of the attributes in this algorithm.

Table 2.2: Tabulation of the SVM algorithm's attribute and the significance of the attributes

No.	Attributes	Significance
1	Internet Protocol (IP) address	The website is phasing if the IP address is used in domain name.
2	URL length	URL length that is more than 75 characters are considered as phishing websites.
3	Shortening service	Link shortened could confusing the user.
4	Having '@' symbol	Websites that contains '@' symbols are usually a phishing website.
5	Double slash redirecting	The website can be categorized as phishing website if there is '/1' at the end of its address.
6	Having sub domain	Websites having more than 2 level and having more than 3 dots (domain within a domain), it could be phishing websites.
7	URL of Anchor	Phishing websites usually have different domains compared to legitimate website, where the anchor tag is connected to the same domain as the source code.

8	Links in tags	It will lead to some infected websites.
9	Abnormal URL	It will extract from the database, and the main identity of the legitimate websites is in the URL.
10	Age of domain	Websites more than six months age can be classified as phishing websites.
11	Page rank	Phishing websites does have low page rank.
12	Links pointing to page	Phishing websites usually have links pointing to zip files that contains malware, which will be downloaded automatically to the computer.

Source : P. Patil et al. (2017)

2.3.4 Comparison among three existing algorithms

Based on the three algorithms above, we can compare among them in terms of their advantages and disadvantages. Table 2.3 below shows the tabulation of the three algorithms including their advantages and disadvantages.

Table 2.3: Tabulation of the three algorithms to prevent the phishing attack.

No.	Algorithm	Advantages	Disadvantages	Example Work
1	Naïve Bayes algorithm	• Has the ability to handle missing values by assimilating the overall opportunities of the missing values (Soofi & Awan, 2017).	• The NB requires a large space to store data due to its instance-based nature where the NB stores all training samples in its process (Archana & Elangovan, 2014)	• Email Spam Classification Using Naïve Bayesian Classifier (Sao & Prashanthi, 2015). • Email Classification using Classification Method (Yitagesu & Tijare, 2016).
		• Its simplicity and quick convergence. It is also an easy and straightforward method (Yasin & Abuhasan, 2016)	• The NB cannot learn about the interactions and relationships between the features in each sample, where it leads to the low accuracy (Yasin & Abuhasan, 2016).	

2	Decision tree algorithm	• The model produced by Decision Tree is easy to be interpreted and understood because it produces simple IF–THEN statements. (Kim, 2016) • The Decision Tree is easy to be implemented compared to others. (Novaković, Strbac & Bulatović, 2011)	• The classification result of Decision Tree is low compared to another ML methods. (Soofi & Awan, 2017) • The Decision Tree is unable to deal with missing values. (Ghamisi, Plaza, Chen, Li & Plaza, 2017)	• Phishing Website Detection using C4.5 Decision Tree (Yang et al., 2017)
3	Support Vector Machine (SVM) algorithm	• The SVM is known for having higher accuracy in classification and its ability to classify data that is not linearly separable. (Yasin & Abuhasan, 2016)	• Hard to implement and handle the numerical variables in the classification problem. (Kim, 2016) • The parameter in SVM is sensitive	• Detecting Spam and Phishing Mails using SVM and Obfuscation URL Detection Algorithm (P. Patil et al., 2017).

		• It is a robust model to solve prediction problems since it maximizes margin. (Byun & Lee, 2002)	where it needs to be set correctly and will affect the classification accuracy if not set properly. (Soofi & Awan, 2017)

2.4 Technologies

Based on the three types of existing algorithm that have been explained above, there are some technologies are used for the algorithm to be implemented. The algorithm can be implemented by using several programming languages, such as Java programming language, C# programming language and Python programming languages. Each of the programming languages have their own advantages and disadvantages. Table 2.4 below shows the tabulation of the types of programming languages including their advantages and disadvantages respectively.

Table 2.4: Tabulation of the types of programming languages.

No.	Programming language	Advantages	Disadvantages
1	C# programming language	• The programming language is a compiled language, which does not allowed the hacker to have access to the source code (Lysis, 2017).	• It is only suitable in the Windows OS environment (Lysis, 2017).

2	Java programming language	• Simple programming language compared to other C derivatives programming languages (Bird, 2015).	• The programming language is usually reference to something else in the program, which sometimes cause the reference problem during runtime (Bird, 2015).
3	Python programming language	• Large number of resources are available in this programming language (Sahouane, 2016).	• Slow and not a good choice of programming language in terms of memory intensive tasks (Sahouane, 2016).

2.5 Conclusion

The comparison among the three types of algorithm has been discussed thoroughly above in terms of their performances, advantages and disadvantages. From the discussion, it shows that each of the algorithm has its own advantages, disadvantages and performances. However, all the algorithms above can be used in preventing the spam email and the phishing attack.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter will focus on the phishing attack detection by using Naïve Bayes algorithm, Decision Tree algorithm and Support Vector Machine algorithm. The dataset that will be used in this study consists of two datasets, which is the email message and the SMS message dataset. This chapter will discuss the overall about the methodology that includes literature study, data collection, classification and performance measurement. The discussion then followed by the timeline of the process related for this study. Besides, it will also include on the hardware and software that will be used in conducting this study.

3.2 Methodology

The research methodology of this study contains four fundamental steps. The steps that related in the research methodology is shown in Figure 3.1. The research methodology starts with literature study, data collection, classification and performance measurement. Next, the discussion of each steps in the research methodology is presented.

3.2.1 Literature review

Based on the discussion on the previous chapter, there are three algorithms that can be used to detect the phishing attack, which is Naïve Bayes algorithm, Support Vector Machine algorithm and Decision tree algorithm. The classification of the datasets helps in detection of the spam content in the messages, which is usually contains phishing websites in it. The comparison between these three algorithms has been shown in previous chapter, which is in Chapter 2. Based on the comparison that made, each method

has their own advantages and disadvantages. The problem statement and objective of this study also has been discussed in Chapter 1.

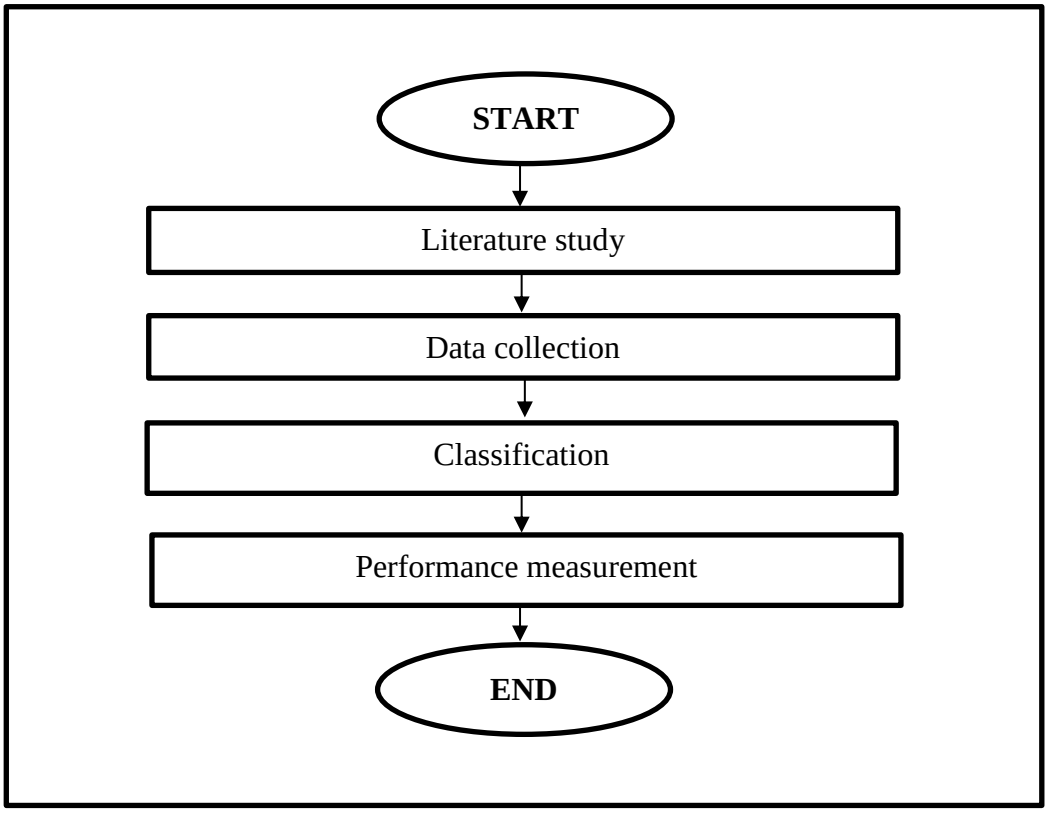


Figure 3.1: The steps in the research methodology

3.2.2 Data collection

In this section, the collection of the data is made. The dataset that will be used in this study is the email messages and the SMS messages, which consists of spam and legitimate (ham) messages. This dataset is used for the training purposes (training set). The information of the dataset that obtained is discuss as below;

3.2.2.1 Dataset 1 – The email Message

The dataset 1 is the email messages that obtained from the GitHub website https://github.com/waleedalinizami/Spam-Detection-Using-Weka (Ali, 2017). The dataset contains 5180 instances and 2 attributes. According to Ali (2017), this dataset features contains a word or character what was frequently occurring in the email. The run-length attributes (55-57) in the email content measure the length of sequences of consecutive capital letters. The format of the dataset is in .ARFF file.

3.2.2.2 Dataset 2 – The SMS Message

The dataset 2 is the SMS messages that obtained from the Unicamp website http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/ (Almeida & José María Gómez Hidalgo, 2011). The dataset contains 5574 instances and 2 attributes. According to Almeida & José María Gómez Hidalgo (2011), this dataset was collected from various sources, such as Grumbletext Web, NUS SMS Corpus and SMS Spam Corpus v.0.1 Big. The format of the dataset is in .ARFF file.

The dataset will be pre-processed before it is ready to be classified. The process will include tokenizing. Tokenising is also known as lexical analysis, which involves in dividing the content of the text into the strings of the character, which known as tokens. Next, the data will be also converted from string to the word vector. According to Jayesh Bapu Ahire (2016), word vector is the vectors of numbers that represent the meaning of a word. Simply to said that the vector of numbers will represent each of the word for the message content in the dataset. The filtering techniques will be also implemented in this step, where the removal of symbols and white space is performed in this phase. After all the steps of data pre-processing is completed, the classification is ready to be done.

3.2.3 Classification

The classification of the word is now performed by using the Bayesian classifier (Naïve Bayes algorithm). This classification is to determine whether the messages is spamming or legitimate (ham) messages. This will use the testing result, which is the number of word frequency that detected as spam words and non-spam words. The Bayes classifier can be calculated by using the Equation (3.1) below;

$$\text{P(Spam/Word)} = \frac{\text{P(Word/Spam) P(Spam)}}{\text{P(Word)}} \quad (3.1)$$

where;

P(Word) = P(Spam) P(Word/Spam) + P(NonSpam) P(Word/NonSpam)

P(Spam/Word) : Probability that an email has word given the email is spam.

P(Word/Spam) : Probability that the word appears in spam email.

P(Word/NonSpam) : Probability that the word appears in legitimate email.

P(Spam) : probability that any given email is spam.

P(NonSpam) : Probability that any given email is non-spam.

The data set will be tested based on the word probability that found in the data set. This test will analyse the word in the messages, whereby the word that found inside the messages will be classify as spam word and non-spam word. It also will be considering the word frequency found in the data set. Once the amount of the word frequency is obtained, then the words can be classified by using Bayesian classifier. This classification process can be done by using WEKA software or using the Java API.

Next, the classification of the word is continued by using the different algorithm, which is Decision Tree algorithm. This algorithm can be expressed in the equation (3.2), (3.3), (3.4), (3.5) and (3.6);

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 \left(p_i \right) \quad (3.2)$$

$$Info_A \left(D \right) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info\left(D_j \right) \quad (3.3)$$

$$SplitInfo_A \left(D \right) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (3.4)$$

$$Gain\left(A \right) = Info\left(D \right) - Info_A \left(D \right) \quad (3.5)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \quad (3.6)$$

where;

D : training set of class-labelled tuple

D_j : subset of D

C_i : the class label of tupple (for i = 1, …, m)

p_i : probability that a tuple in D belongs to class C_i

|D| : the number of tuples in D

Basically, this algorithm will classify the words by generating the decision tree. Each node of the tree, the algorithm will choose an attribute in the data that can further split the samples into subsets. Each of the leaf node will represents a classification or decision. There are also some premises guide on this algorithm, which described below;

a) The tree is a leaf if all the cases are of the same class.
b) The potential information that provide by a test on the attribute and the gain in information that would result from a test on the attribute is calculated for each attribute.
c) The best attribute to branch on will be find depending on the current selection criterion.

Besides, Support Vector Machine algorithm will be used on the classification of the word. This algorithm can be expressed in the equation (3.7) and (3.8);

$$\min \frac{1}{2} |w|^2 + c \sum_{i=1}^{n} \xi_i \tag{3.7}$$

$$y_i \left(w x_i - b \right) >> 1 - \xi_i \qquad \xi_i \geq 0 \tag{3.8}$$

where;

i = 1, 2, …, n

n : the dimensionality of the feature

x : input vector

w : the normal vector to the hyperplane

C : capacity constant

ξ_i : parameters for handling no separable data (inputs)

This algorithm also known as a linear strong classifier, which would able to identify two classes label in a dataset. It will produce a set of hyperplanes and the maximum marginal hyperplane will be considered at the end of the experiment. Based on this study, the maximum marginal hyperplane may contain the spam or legitimate (ham) messages.

Cross-validation method will be used in this study. The purpose of using this method is that to improve the accuracy of the classification of the datasets. There are many types of cross-validation process, which includes *k*-fold cross-validation and leave-*p*-out cross-validation. The one that will be used is *k*-fold cross-validation. According to Gupta (2017), *k*-fold cross-validation is the process where the data will be divided into *k* subsets. Based on the *k* subsets, in each time, one of the *k* subsets is used as the test set and the other *k*-1 subsets are put together to form a training set. This will cut down bias as the classification are using most of the data for fitting. Besides, it will also cut down the variances as most of the data is also being used in test set. All the results that obtained will be recorded and the average of the result will be calculate.

3.2.4 Performance measurement

Based on the result that obtained, the performance measurement can be measured and evaluated. This performance measurement includes the time taken for the training time, the correctly classified instances, the percentage of correctly classified instances (or known as the accuracy percentage), the True Positive rate, False Positive rate and the precision rate. The accuracy of the result is the ability of the algorithm to classify the instances correctly. The precision rate of the result refers to how close the estimates from different samples are to each other. The both dataset classification results that obtained will be compared based on the performance measurement that stated above. The accuracy and precision can be calculated based on the Equation (3.2) and (3.3) below;

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+TP+FN+FP}} \tag{3.9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3.1}$$

where;

TN : true negative (legitimate predicted as legitimate)

TP : true positive (spam predicted as spam)

FP : legitimate predicted as spam

FN : spam predicted as legitimate

3.3 Hardware and software

There are some hardware and software that will be used for the phishing attack detection by using Naïve Bayes algorithm. Table 3.1 and 3.2 shows the hardware and software requirement and specification that will be used in this method for the training purpose respectively.

Table 3.1: The hardware requirements and specification

Hardware	**Specification**
Acer Aspire E-15 (E5-575G-58F1) Notebook PC	Processor Intel Core™ i5 (7th Generation) Windows 10 Home Single Language

Table 3.2: The software requirements and specification

Hardware	**Specification**
Microsoft Office Word 2016	Used to write the research report from Chapter 1 to Chapter 5.
Mendeley Desktop	Used to generate the citation of the articles, research paper etc. that used in this research.
MathType 6.9b	Used to generate the equation in the research report.
Microsoft Project Professional 2016	Used to construct the Gantt Chart of this research.
NetBeans IDE 8.2	Used to compile the Java code and run the test of the dataset.

3.4 Gantt Chart

In this section, the Gantt chart is created to show the suggested progress for the study. This is to ensure that the study can be conducted in an organized and proper way. Figure 3.2 to Figure 3.4 shows the Gantt chart of this study.

ID	Task Mode	Task Name	Duration	Start	Finish
1		**Phase 1: Defining problems**	**10 days**	**Mon 12/2/18**	**Fri 23/2/18**
2		Study the phishing and spam attack	6 days	Mon 12/2/18	Mon 19/2/18
3		Identifying the problems, goals, objective, scope and significance of the study	4 days	Tue 20/2/18	Fri 23/2/18
4		Documenting the Defining Problems	3 days	Mon 26/2/18	Wed 28/2/18
5		**Phase 2: Literature Review**	**20 days**	**Thu 1/3/18**	**Wed 28/3/18**
6		Research on Machine Learning methods, which is Naïve Bayes algorithm, Decision Tree algorithm and Support Vector Machine	12 days	Thu 1/3/18	Fri 16/3/18
7		Analyze and compare the advantages	3 days	Mon 19/3/18	Wed 21/3/18
8		Documenting the Literature Review	5 days	Thu 22/3/18	Wed 28/3/18

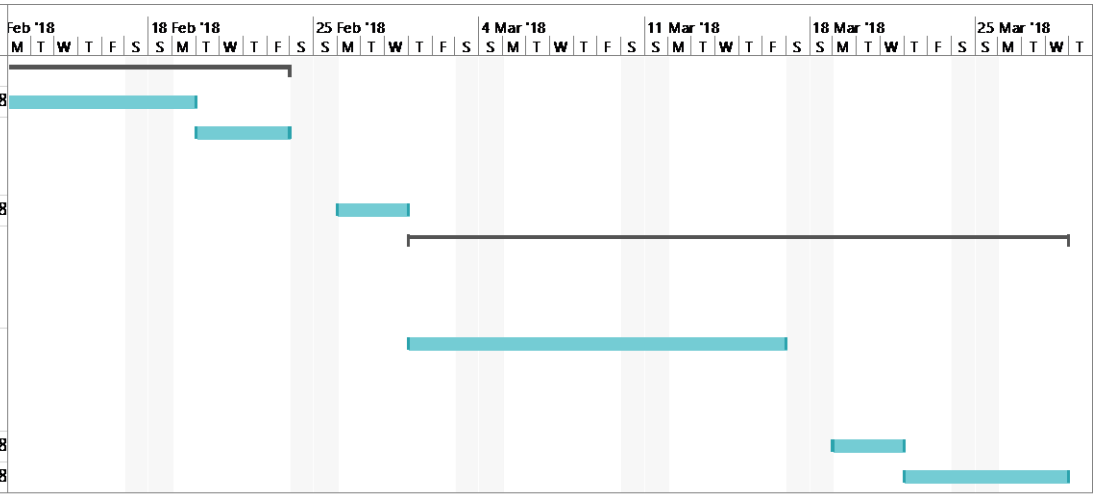


Figure 3.2: The Gantt chart from phase 1 to phase 2.

ID	Task Mode	Task Name	Duration	Start	Finish
9		**Phase 3: Methodology**	**12 days**	**Thu 29/3/18**	**Fri 13/4/18**
10		Study on the Research Methodology	2 days	Thu 29/3/18	Fri 30/3/18
11		Create Research Methodology based on the study	5 days	Mon 2/4/18	Fri 6/4/18
12		Hardware and software specifications	1 day	Mon 9/4/18	Mon 9/4/18
13		Gantt Chart	1 day	Tue 10/4/18	Tue 10/4/18
14		Documenting the Methodology	3 days	Wed 11/4/18	Fri 13/4/18
15		**Phase 4: Implementation and Testing**	**164 days**	**Mon 16/4/18**	**Thu 29/11/18**
16		Testing the Java coding implementation	100 days	Mon 16/4/18	Fri 31/8/18

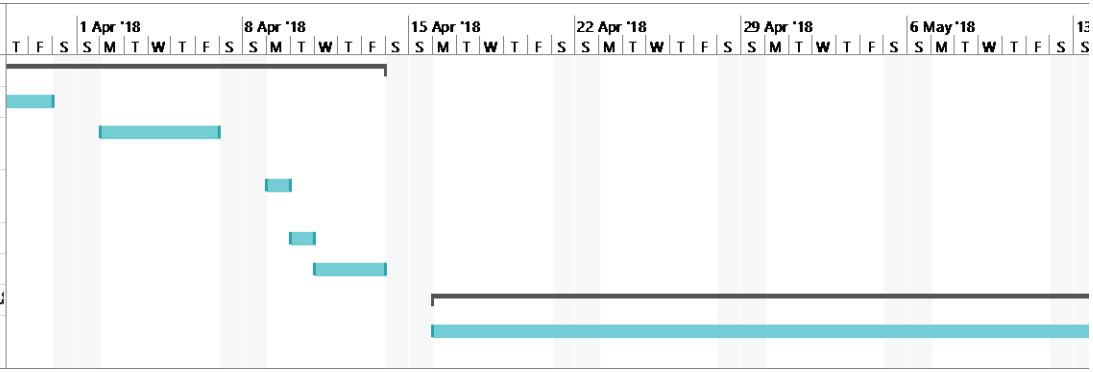


Figure 3.3: The Gantt chart from phase 3 to phase 4.

ID	Task Mode	Task Name	Duration	Start	Finish
17		Finding the dataset to be used	16 days	Mon 3/9/18	Mon 24/9/18
18		Run the experiment on the dataset	12 days	Tue 25/9/18	Wed 10/10/1
19		Record and tabulate the results that obtained	4 days	Thu 11/10/18	Tue 16/10/18
20		Discussion on the results based on the experiment	22 days	Wed 17/10/18	Thu 15/11/18
21		Documenting the Implementation and Testing	10 days	Fri 16/11/18	Thu 29/11/18
22		**Phase 5: Documentation**	**7 days**	**Fri 30/11/18**	**Mon 10/12/1**
23		Final documentation	7 days	Fri 30/11/18	Mon 10/12/1

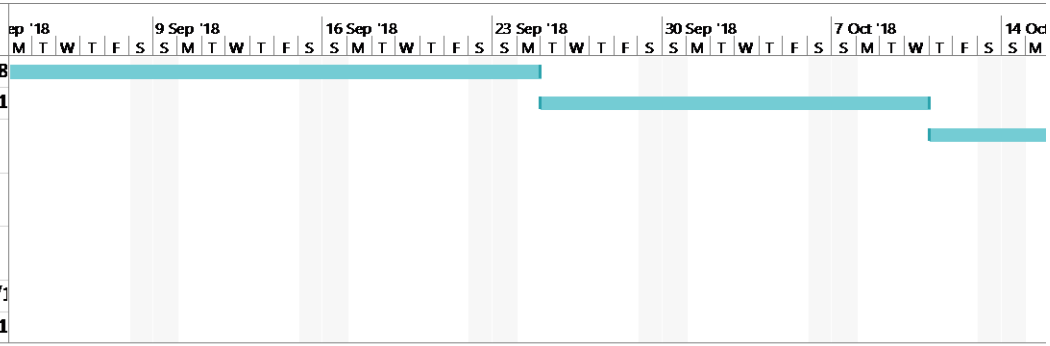


Figure 3.4: The Gantt chart from phase 4 to phase 5.

3.5 Implementation

The implementation of this study starts with the dataset collection. The dataset that will be use in this study is the email messages and the SMS messages, which contains of spam and legitimate (ham) messages. This dataset may contain the phishing website, which lead to the phishing attack. The first dataset (the email message dataset) contain 5180 instances and 2 attributes, whereas the second dataset (the SMS message dataset) contain 5574 instances and 2 attributes. After collecting the dataset, the cross-validation method is used to split out the dataset into 10 parts. The cross-validation method will be used is 10-fold cross-validation method.

Next, the classification can be done by running the experiment each of the dataset that has been split out. The algorithm that will be used is Naïve Bayes algorithm. The Decision tree algorithm and the Support Vector Machine algorithm will be also used in classifying the datasets. The experiment will be run 10 times for each of the dataset. The result of each of the split dataset will be recorded. The experiment starts by running the Java coding in NetBeans.

The process of the Java coding is that it will read the dataset, which is in .ARFF format. The dataset that is successfully read will then go through the pre-processing part. This part includes the filtering process of the data, which is eliminating the symbols and white space in the dataset. Besides, the word in the message will be converted to word vector. This process is one of the process of tokenization, where the word vector will represent the words in the dataset. The classification can be made once the pre-processing part of the dataset are finished. The dataset will be classified by using the algorithm that is set.

After the experiment, the average reading of the result will be used to represent the result for each dataset. The criteria that will be collected includes the time taken for the training time, the correctly classified instances, the percentage of correctly classified instances (or known as the accuracy percentage), the True Positive rate, False Positive rate and the precision rate. This will be done for each of the dataset, which is the email message dataset and SMS message dataset. The reading of each the criteria that stated above will be displayed once the dataset has successfully classified. Figure 3.3 below shows the summarization of the steps that taken in the implementation phase.

Basically, the higher the value of accuracy, the more accurate the results be. This study suggested that the combination with the different algorithm can be made so that the accuracy of detecting the spamming messages can be improved. Besides, minimizing the number of instances in the test set could be help in improving the accuracy of the classification.

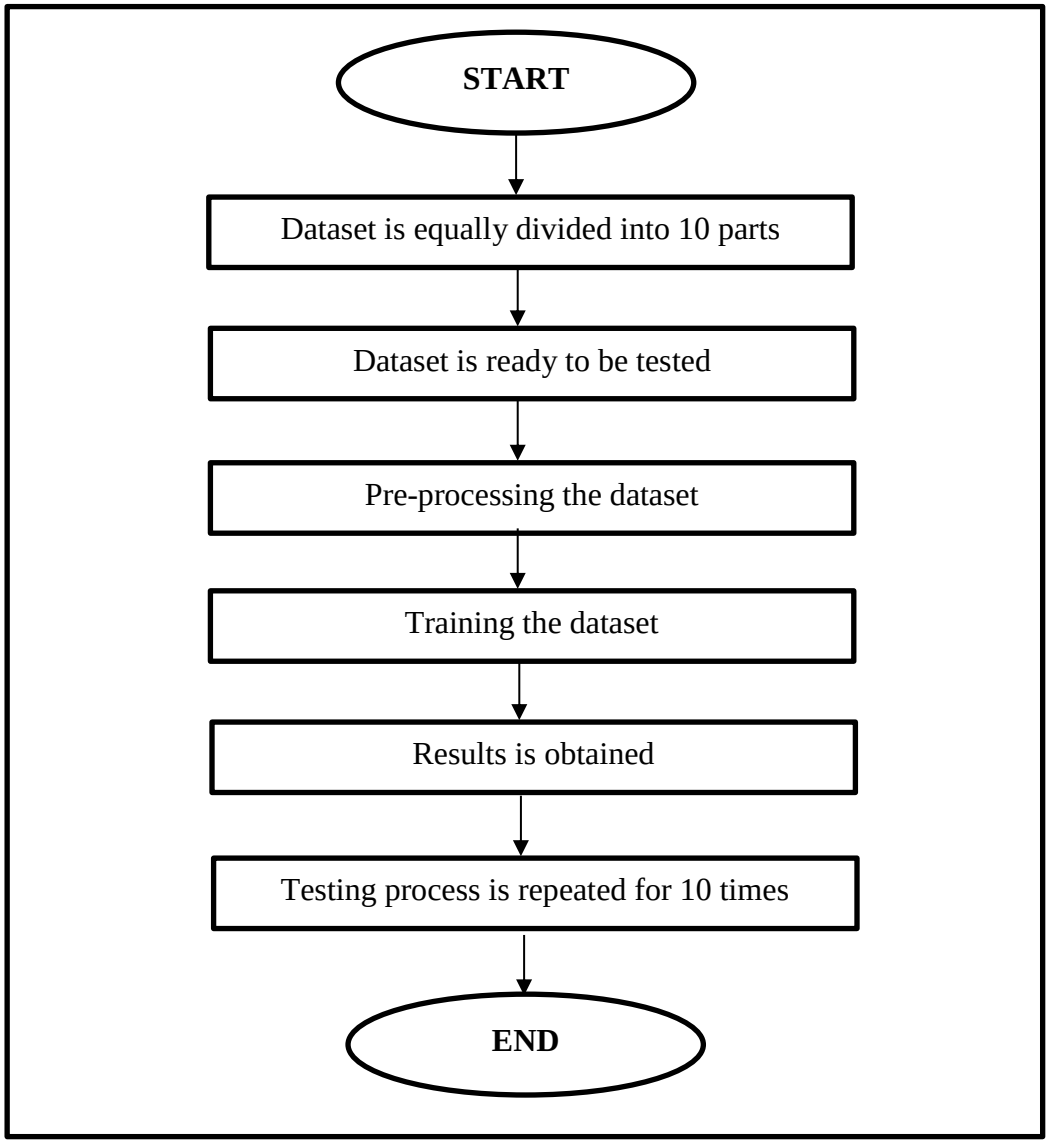


Figure 3.5: The summary of the implementation process.

3.6 Conclusion

Based on the discussion above, the research is focus on the phishing attack detection by using machine learning method. This algorithm consists of four fundamental steps, which includes the literature study, data collection, classification and performance measurement. Each of the steps have been discussed thoroughly. Besides, the Gantt chart also provided, showing the suggested progress of the study to be conducted. Finally, the implementation of this study, which discuss on how the classification of the dataset is perform by using Java coding in NetBeans, is also included in this chapter.

CHAPTER 4

TESTING AND RESULT DISCUSSION

4.1 Introduction

This chapter will discuss on the implementation of the spam detection by using machine learning method, which is Naïve Bayes algorithm, Decision tree algorithm and Support Vector Machine algorithm. The testing involves two datasets, which consists of email message dataset and SMS message dataset. The results are produced by conducting the experiment. These results will be collected and compared between the two datasets. The results are tabulated so that comparison can be made.

4.2 Results

The experiment was conducted by using machine learning methods, which were Naïve Bayes algorithm, Decision tree algorithm and Support Vector Machine algorithm. After doing the experiment, the results are obtained for both datasets. It shows different reading on the several aspects such as the number of correctly classified instances and the time taken for the experiment conducted.

4.2.1 Dataset 1: The email messages

The dataset 1, which is the email message, has been experimented by using the both algorithms. The results are obtained and tabulated. Table 4.1 shows the tabulation of the result that obtained.

Table 4.1: Tabulation of the result of dataset 1

Evaluation Criteria / Classifier	Training time (minutes)	Correctly classified instances (out of 5180)	Average percentage of correctly classified instances (%)	TP Rate (Average)	FP Rate (Average)	Precision
Naïve Bayes algorithm	01:23	4990	95.6646	0.964	0.037	0.965
Decision Tree algorithm	01:45	5079	98.0502	0.980	0.019	0.981
Support Vector Machine algorithm	00.13	5180	100	1	0	1

Based on the result, the training time takes longer on the Decision tree algorithm compared to the Naïve Bayes algorithm. Besides, the total number and the percentage of instances that correctly classified shows Decision tree algorithm has better performance if compared to Naïve Bayes algorithm. This trend is also followed by the TP Rate and the precision reading. Since the Decision tree algorithm shows higher TP rate, the FP rate of this approach is smaller if compared to the Naïve Bayes algorithm. However, there is 100% correctly classified instances if the Support Vector Machine algorithm is used. The time taken for the classification in this algorithm also shows the shortest among them. Figure 4.1 below shows the bar chart of the percentage based on the results;

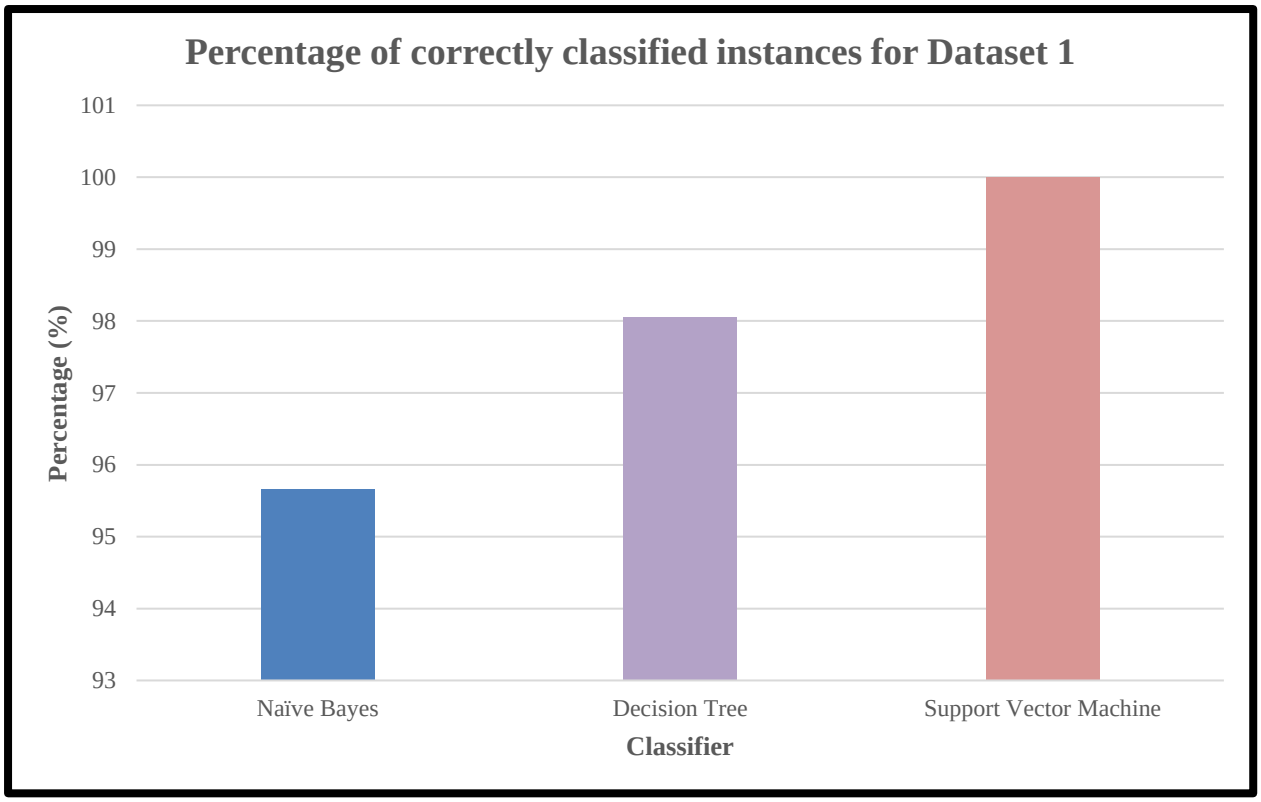


Figure 4.1: Graph of percentage of correctly classified instances based on Dataset 1

Based on the result that obtained above, overall, it shows that the dataset that tested is better when using the Support Vector Machine algorithm compared to Decision tree algorithm and Naïve Bayes algorithm.

4.2.2 Dataset 2: The SMS messages

The dataset 2, which is the SMS message has been experimented by using the both algorithms. The results are obtained. Table 4.2 shows the tabulation of the result that obtained.

Table 4.2: Tabulation of the result of dataset 2

Evaluation Criteria / **Classifier**	**Training time (minutes)**	**Correctly classified instances (out of 5574)**	**Average percentage of correctly classified instances (%)**	**TP Rate (Average)**	**FP Rate (Average)**	**Precision**
Naïve Bayes algorithm	00.10	5515	98.9415	0.990	0.041	0.989
Decision Tree algorithm	00.18	5395	96.7887	0.967	0.172	0.967
Support Vector Machine algorithm	00.07	5573	99.8205	0.998	0.010	0.998

Based on the result above, the training time takes longer on the Decision Tree algorithm compared to the Naïve Bayes algorithm. However, the total number and the percentage of instances that correctly classified shows Naïve Bayes algorithm has better performance if compared to Decision tree algorithm. This trend is also followed by the TP Rate and the precision reading. Since the Naïve Bayes algorithm shows higher TP rate, the FP rate of this approach is smaller if compared to the Decision tree algorithm. The Support Vector Machine algorithm shows that it has the shortest time taken among them. The percentage of the instances that correctly classified also shows that it is the highest percentage among them. Figure 4.2 below shows the bar chart of the percentage based on the results;

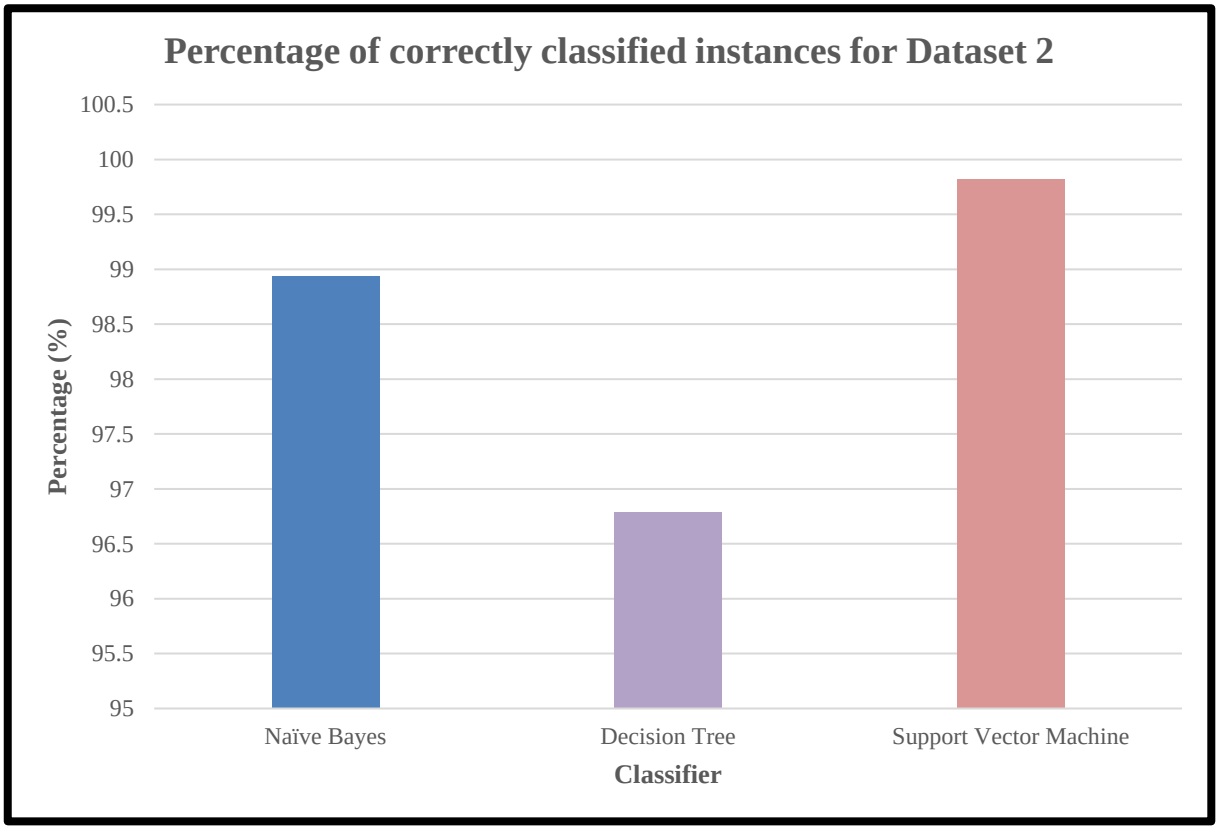


Figure 4.2: Graph of percentage of correctly classified instances based on Dataset 2

Based on the result that obtained above, overall, it shows that the dataset that tested is better when using the Naïve Bayes algorithm compared to Decision tree algorithm.

4.3 Discussion

The results that obtained are interpreted and compared. From the results, in this experiment, it shows that different dataset has different performance level and accuracy level. For example, based on the dataset 1, the dataset takes longer time for the classifier to classify the dataset whether it is a spam, ham or norm. However, there is only shorter time for the dataset 2 to be tested using this both algorithms. This is because the content of the dataset 1 is huge in the number of words of each instance, whereas dataset 2 only contains small number of words of each instance. This will cause the classifier to take time much longer in analysing the content of the words in the dataset 1 compared to dataset 2. This statement also supported by Mack (2018), where the size of the dataset and the number of attributes in a data will affects the time taken for the classification to complete. Figure 4.3 and Figure 4.4 shows the graph showing the time taken for the both datasets that classified by using Naïve Bayes algorithm, Decision tree algorithm and Support Vector Machine algorithm respectively.

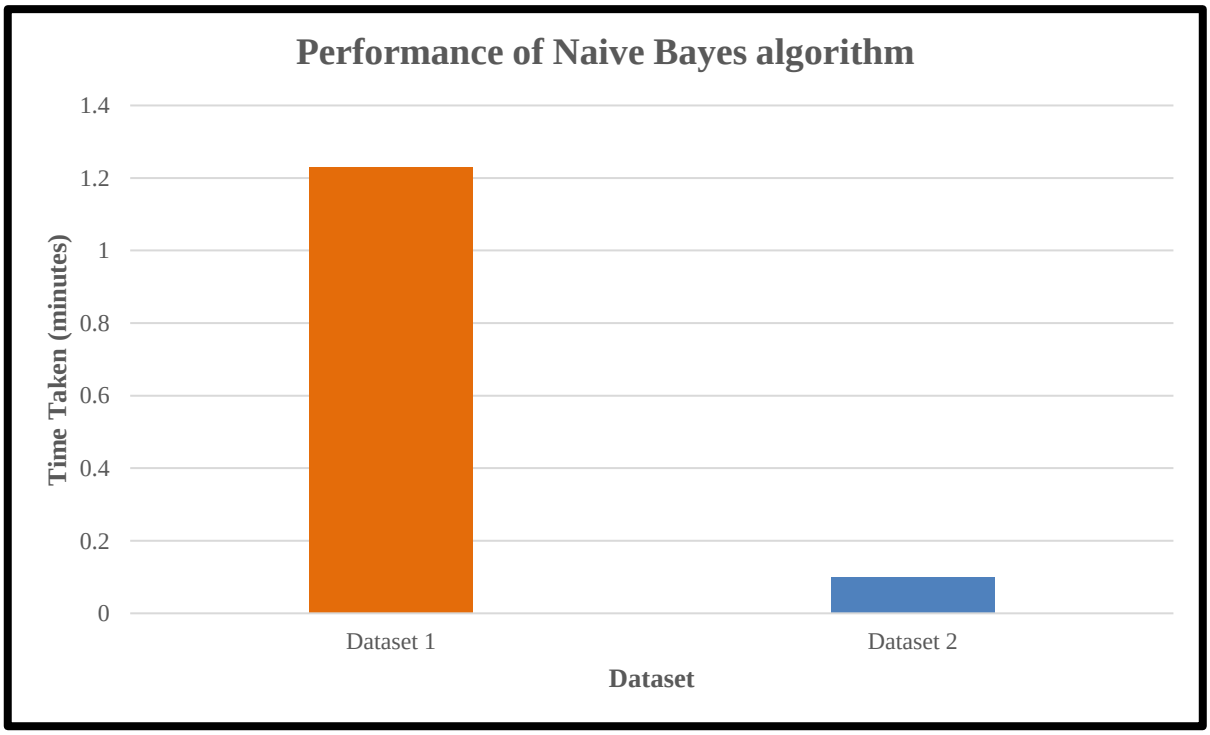


Figure 4.3: Graph of time taken for the classification of dataset using Naïve Bayes algorithm.

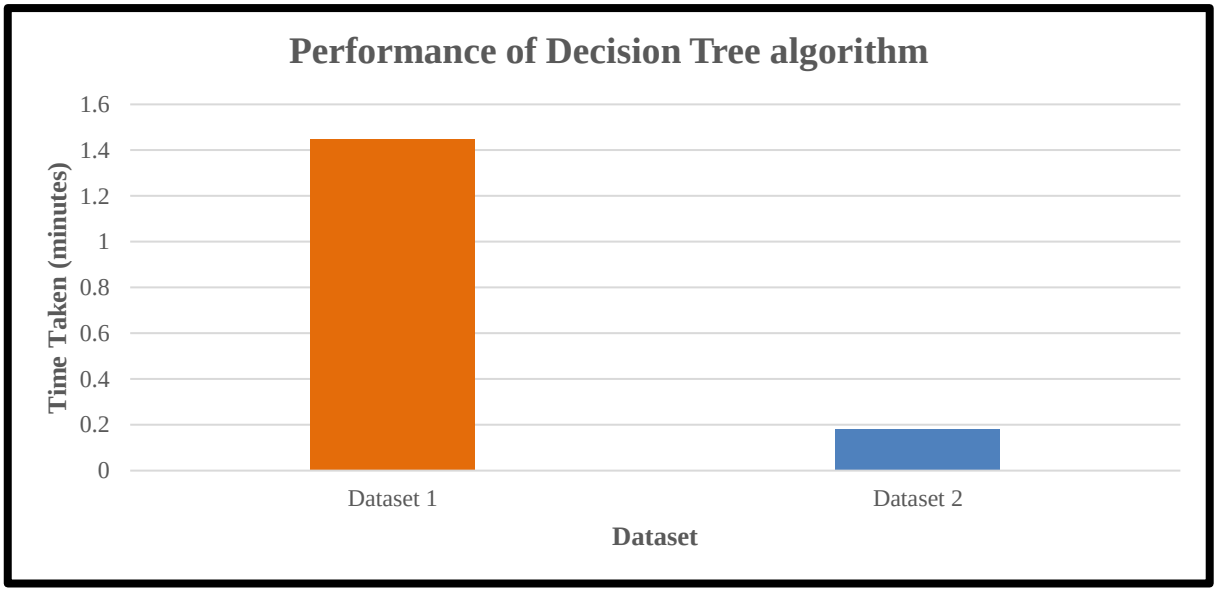


Figure 4.4: Graph of time taken for the classification of dataset using Decision Tree algorithm.

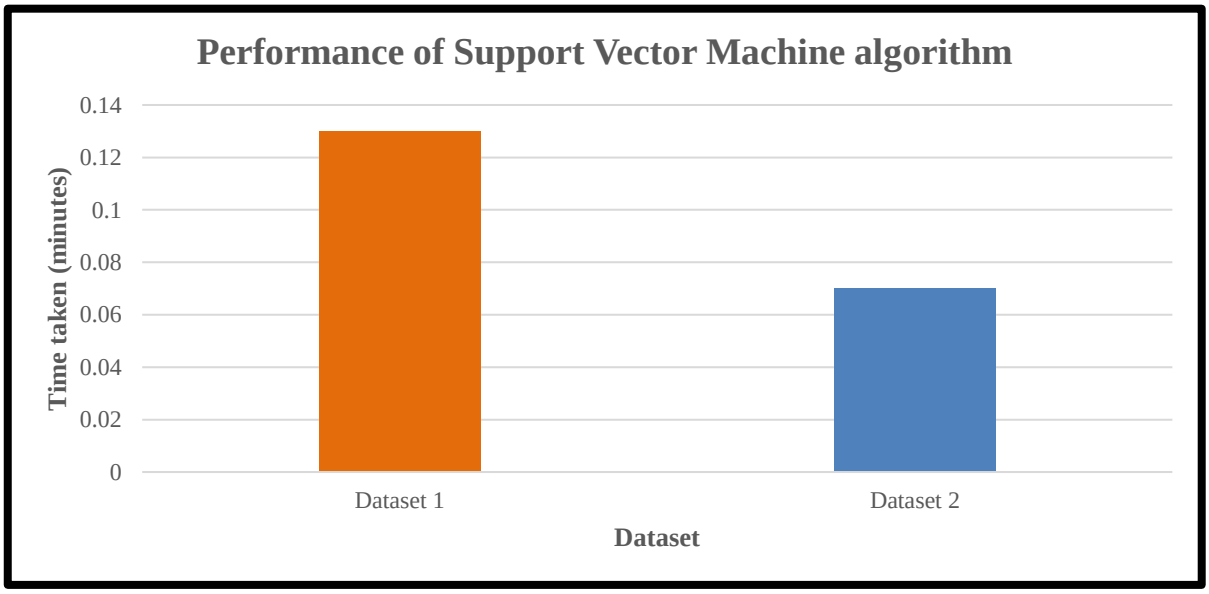


Figure 4.5: Graph of time taken for the classification of dataset using Support Vector Machine algorithm.

For the performance based on the previous research, each of the datasets has a different performance level if compared to this study. Table 4.3 and 4.4 shows the tabulation of the results for dataset 1 and 2 based on the previous research.

Table 4.3: Tabulation of the results based on the previous research (Metsis, Androutsopoulos, & Paliouras, 2006)

Naïve Bayes version	**Average percentage correctly classified instances (%)**
Flexible Bayes	95.99
Multi-variate Gauss NB	92.32
Multinomial NB (TF attribute)	97.13
Multi-variate Bernoulli NB	96.52
Multinomial NB (Boolean attribute)	97.53

Table 4.4: Tabulation of the results based on the previous research (Almeida, Hidalgo, & Yamakami, 2011)

Classifier	**Average percentage correctly classified instances (%)**
Naïve Bayes algorithm	92.70
Decision tree algorithm	95.00
Support Vector Machine algorithm	97.64

Based on the previous research results in Table 4.3, the study uses different types of Naïve Bayes algorithm approach by using dataset 1. The results of each approach will be different. If compared to this study, in dataset 1, the Naïve Bayes algorithm achieves 95.6646% of the correctly classified instances. There is a slightly difference if compared to the previous research results. This is due to the approaches that used on the Naïve Bayes algorithm itself. We can say that different approaches had different performance level.

Besides, in dataset 2, it shows that the Naïve Bayes algorithm shows better performances compared to the Decision Tree algorithm. However, if the results obtained in this study are compared with the previous research (Almeida et al., 2011) based on Table 4.4, it was found out that the Decision Tree algorithm achieve the high accuracy compared to Naïve Bayes algorithm. This is because the previous research is using combination of the algorithm with the tokenization to improve the accuracy result of the classification. The tokenization that used is that it considers the separation of the words in the dataset, such as blanks, tabs and commas as the tokens. This will preserve other symbols that may help to separate spam and legitimate message. It can be simply said that the tokenization approach that used in the previous research is different compared to this study. However, the level of performance for the Support Vector Machine algorithm is maintained compared to the other two algorithms, which achieve the highest percentage of the correctly classified instances.

Next, the use of filter for the dataset pre-processing, which is StringtoWordVector filter may also contribute to the accuracy reading for the classification. StringtoWordVector is the assumption that made where the document text in an attribute of type String is a nominal attribute (Witten, Frank, Hall, & Pal, 2016). It is a kind of filter that found in WEKA, where it is simply makes the attribute value in the transformed dataset 1 or 0 for all single-word terms, depending on whether the word appears in the document or not. The assumption on the words in the dataset may cause the accuracy result to be different from the previous research. This can be shown if the experiment is conducted on dataset 1 when there is no StringtoWordVector filter in it by using WEKA software. The classifier that used will be the Multinomial Naïve Bayes. It shows the reading that only 2.5% can be correctly classified.

4.4 Conclusion

Based on the experiment that conducted, it shows that different algorithm had different performance level and accuracy. However, there are several aspects that need to be consider in conducting the experiment. This includes the use of filter in preparing the datasets or can be known as pre-processing the datasets and the amount of word count of each instances in the dataset. This may also contribute to the performance level and the accuracy of the classification.

CHAPTER 5

CONCLUSION

5.1 Concluding remarks

Machine learning method is one of the methods in detection of the phishing attack as well as the spam attack. In this research, three algorithms in the machine learning method are used, which is Naïve Bayes algorithm, Decision Tree algorithm and Support Vector Machine algorithm. The detection will classify the word content in the dataset. From the result that obtained, it shows the performance of the algorithm in each of the dataset. The performance criteria includes the time taken for the training time, the correctly classified instances, the percentage of correctly classified instances (or known as the accuracy percentage), the True Positive rate, False Positive rate and the precision rate.

There are four main steps that related in this study. The step includes literature study, data collection, classification and the performance measurement. The dataset is trained by using the three of the algorithms, which is done by running the Java coding in NetBeans. Before the classification starts, the data is pre-processed in order to get an accurate result. The words that contained in the messages will be converted into string vector. This step uses StringtoWordVector filter, which categorized as unsupervised filter that can found in the WEKA classifier. The Java API for WEKA is installed for the filter to be used in the Java coding. The classification started once the data pre-processing finished. The results of the classification are generated. The results than will be compared and discussed.

5.2 Research constraints and challenges

There are few constraints faced during the experiment conducted. First, the accuracy level obtained in this study is different compared to the previous research. This is because the different method of data pre-processing method that used in this study. Besides, the longer time taken needed for the classifier to classify the words in the dataset. This may affect the performance level of the algorithm that used in classification of the datasets. The timing constraints also need to be taken in consideration. There is a limited time to study the performance of the classifier, which is the algorithms that used.

5.3 Future work

The results that obtained are interpreted and compared. From the results, in this experiment, different method used in pre-processing the dataset will affect to the accuracy level of the classification. This need to be consider in the future time so that the level of the accuracy of the classification can be maintained and optimized. This study suggests that the classification of the spam message can be go through in detail so that the phishing attack can be prevented in the network environment, especially in huge organization.

REFERENCES

A. Yasin and A. Abuhasan, "An Intelligent Classification Model For Phishing Email Detection," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 8, no. 4, 2016.

Aized Amin Soofi and Arshad Awan, "Classification Techniques in Machine Learning: Applications and Issues," *Journal of Basic & Applied Sciences*, vol. 13, pp. 459–465, 2017.

Akansha, P., & Meenakshi, E. (2017). Detection of Phishing Websites Using Data Mining Techniques, *2*(12), 1468–1472.

Ali, W. (2017). Spam Detection using WEKA. Retrieved from https://github.com/waleedalinizami/Spam-Detection-Using-Weka

Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering. *Proceedings of the 11th ACM Symposium on Document Engineering - DocEng '11*, 259. https://doi.org/10.1145/2034691.2034742

Almeida, T. A., & José María Gómez Hidalgo. (2011). SMS Spam Collection v.1. Retrieved from http://dcomp.sor.ufscar.br/talmeida/smspamcollection/

Bird, M. (2015). The Pros and Cons of using Java. Retrieved March 22, 2018, from http://www.digitalrise.biz/software/the-pros-and-cons-of-using-java/

Gupta, P. (2017). Cross-Validation in Machine Learning. Retrieved from https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

H. Byun and S.-W. Lee, "Applications of Support Vector Machines for Pattern Recognition: A Survey," in *Pattern Recognition with Support Vector Machines*, 2002, pp. 213–236.

Jasmina Novaković, Perica Strbac, and Dusan Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research* , vol. 21, no. 1, pp. 191–135, 2011.

Jayesh Bapu Ahire. (2016). Introduction to Word Vector, 1–11. Retrieved from https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf

K. Kim, “A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree,” *Pattern Recognition*, vol. 60, pp. 157–163, Dec. 2016.

Kaur, S., & Kaur, A. (2015). Detection of Phishing Webpages using Weights computed through Genetic Algorithm. In *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)* (pp. 331–336). Amritsar: IEEE.

Kozak, J., & Boryczka, U. (2016). Collective data mining in the ant colony decision tree approach. *Information Sciences*, *372*, 126–147. https://doi.org/10.1016/j.ins.2016.08.051

Kumar, N., & Chaudhary, P. (2017). Mobile Phishing Detection using Naive Bayesian Algorithm, *17*(7), 142–147.

Lysis. (2017). Pros and Cons of Using C# as Your Backend Programming Language. Retrieved March 22, 2018, from https://agilites.com/pros-and-cons-of-using-c-as-your-backend-programming-language.html

Mack, D. (2018). How to pick the best learning rate for your machine learning project. Retrieved from https://medium.freecodecamp.org/how-to-pick-the-best-learning-rate-for-your-machine-learning-project-9c28865039a8

Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective* (2nd ed.). Boca Raton: Taylor & Francis Group.

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering using Naive Bayes-Which Naive Bayes? https://doi.org/10.3109/02841866309134119

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review*, *17*, 1–24. https://doi.org/10.1016/j.cosrev.2015.04.001

P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, “Advanced Spectral Classifiers for Hyperspectral Images: A review,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8–32, 2017.

Patil, P., Rane, R., & Bhalekar, M. (2017). Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm. *Proceedings of the International Conference on Inventive Systems and Control, ICISC 2017*, 1–4. https://doi.org/10.1109/ICISC.2017.8068633

Patil, T. R. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications, ISSN: 0974-1011*, *6*(2), 256–261. https://doi.org/ISSN: 0974-1011

Pawar, M. V., & Anuradha, J. (2015). Network security and types of attacks in network. *Procedia Computer Science*, *48*(C), 503–506. https://doi.org/10.1016/j.procs.2015.04.126

Rathod, S. B., & Pattewar, T. M. (2015). Content Based Spam Detection in Email using Bayesian Classifier. In *Conference: 2015 International Conference on Communications and Signal Processing (ICCSP)* (pp. 1257–1261). https://doi.org/10.1109/ICCSP.2015.7322709

Ray, S. (2017). Understanding Support Vector Machine algorithm from examples (along with code). Retrieved March 22, 2018, from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

S.Archana and K.Elangovan, "Survey of Classification Techniques in Data Mining," *International Journal of Computer Science and Mobile Applications*, vol. 2, no. 2, pp. 65–71, 2014.

Sahouane, A. (2016). The pros and cons of Python. Retrieved March 22, 2018, from https://www.supinfo.com/articles/single/3425-the-pros-and-cons-of-python

Sao, P., & Prashanthi, P. K. (2015). E-mail Spam Classification Using Naïve Bayesian Classifier. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *4*(6), 2792–2796.

Saxena, R. (2017a). How Decision Tree Algorithm Works. Retrieved March 22, 2018, from http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

Saxena, R. (2017b). SVM Classifier, Introduction to Support Vector Machine Algorithm. Retrieved April 5, 2018, from http://dataaspirant.com/2017/01/13/support-vector-machine-algorithm/

Smadi, S., Aslam, N., Zhang, L., Alasem, R., & Hossain, M. A. (2016). Detection of phishing emails using data mining algorithms. In *SKIMA 2015 - 9th International Conference on Software, Knowledge, Information Management and Applications* (pp. 1–8). https://doi.org/10.1109/SKIMA.2015.7399985

Suganya, V. (2016). A Review on Phishing Attacks and Various Anti Phishing Techniques. *International Journal of Computer Applications*, *139*(1), 975–8887.

Waldron, M. (2015). Naive Bayes for Dummies; A Simple Explanation. Retrieved March 22, 2018, from http://blog.aylien.com/naive-bayes-for-dummies-a-simple-explanation/

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (Fourth). Cambridge: Elsevier Inc.

Yang, X., Yan, L., Yang, B., & Li, Y. (2017). Phishing Website Detection Using C4 . 5 Decision Tree, (Itme), 119–124.

Yitagesu, M. E., & Tijare, P. M. (2016). Email Classification using Classification Method, *32*(3), 142–145.