CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

MONTHIYA SAPAN

A THEMATIC PAPER SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER DEGREE OF SCIENCE (INFORMATION TECHNOLOGY MANAGEMENT) FACULTY OF GRADUATE STUDIES MAHIDOL UNIVERSITY 2017

Thematic Paper
entitled

CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

มนทิยา สาปาน
.....
Miss Monthiya Sapan
Candidate

.....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Major advisor

.....
Lect. Taweesak Samanchuen,
Ph.D. (Electrical Engineering)
Co-advisor

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies,
Mahidol University

Adisorn Leelasantitham
.....
Assoc. Prof. Adisorn Leelasantitham,
Ph.D. (Electrical Engineering)
Acting Program Director
Master of Science Program in
Information Technology Management
Faculty of Engineering
Mahidol University

Thematic Paper
entitled

CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Information Technology Management)
on
January 6, 2017

มณทิยา สาปาน
.....
Miss Monthiya Sapan
Candidate

.....
Lect. Smitti Darakorn Na Ayuthaya,
Ph.D. (Public Administration)
Chair

.....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Member

Sustarum Thammaboosadee
.....
Asst. Prof. Sustarum Thammaboosadee,
Ph.D. (Political Science-International Relations)
Member

.....
Lect. Taweesak Samanchuen,
Ph.D. (Electrical Engineering)
Member

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies,
Mahidol University

.....
Asst. Prof. Jackrit Suthakorn,
Ph.D. (Robotics)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

Thematic Paper has been completed and succeeded due to the guidance and assistance of several people. I would first like to say a very big thank you to my advisor, Lect.Sotarat Thammaboosadee, for his attentive support, thoughtful guidance, and correction of thematic Paper.

I would also like to express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of writing this thematic Paper. My thanks extend to all of my friends for their great encouragement, to all professors and staff for my journey of study at IT Management. Finally, I am so thankful to all ITM58 friends who have supported me through a tough year of study. Thank you

Monthiya Sapan

CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

MONTHIYA SAPAN 5837645 EGIT/M

M.Sc. (INFORMATION TECHNOLOGY MANAGEMENT)

THEMATIC PAPER ADVISORY COMMITTEE: SOTARAT THAMMABOOSADEE, Ph.D., TAWEESAK SAMANCHUEN, Ph.D.,

ABSTRACT

Different types of internet attack have currently increasing exponentially. One of internet attacks that has been used for many years. Currently, Phishing and number of internet users who have been attacked by phishing has also been increasing; this trend has causes a large scale of losses to victims. This research studies contents in phishing email only. Text classification system was applied for analyzing phishing email contents based on the specified eight features, including studying campaign messages that appeared in phishing emails and determining the words used in those messages. The dataset of this study is provided by www.419scam.org and the results were used to create a decision tree. The overall model performance is greater than 80% when binary occurrence is used as an indicator. The decision making rules are further analyzed facilitated by the association rules discovery method to determine the relation of features for creating the final phishing determination model. When analyzing the relationship of features, the relation rule was obtained and the emails that included Messages which notified the recipients that the e-mail is confidential, Messages which rushed the recipients to take an immediate action, and Message which asked for help, were considered is be Phishing E-Mail. This research could help in analyzing email contents and determining whether there is a risk of them being phishing emails. This could be a part of reducing risk of being attacked by email phishing. In the future, it is therefore suggested the research should be extended to analyzing other email components such as the domain reliability and files attached in email.

KEY WORDS: PHISHING EMAIL/ CRAFTING TEXT/ DATA MINING/ TEXT MINING/ DECISION TREE/ MODULAR MODEL

63 pages

แบบจำลองสำหรับจำแนกข้อความหลอกลวงแบบหน่วยย่อยตามเนื้อหาสำหรับการตรวจจับอีเมล์ฟิชชิ่ง

CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

มณฑิยา สาปาน 5837645 EGIT/M

วท.ม. (การจัดการเทคโนโลยีการสารสนเทศ)

คณะกรรมการที่ปรึกษาสารนิพนธ์ : โยฑศ์รัตต์ ธรรมบุษดี, Ph.D., ทวีศักดิ์ สมานชื่น, Ph.D.,

บทคัดย่อ

การโจมตีจากภัยคุกคามต่างๆบนโลกอินเทอร์เน็ตในปัจจุบัน นับว่ามีความรุนแรงที่ทวีคูณมากขึ้นเรื่อยๆ และมีรูปแบบการโจมตีที่หลากหลายมาก Phishing ก็ถือว่าเป็นภัยคุกคามรูปแบบหนึ่งที่มีมานานและในปัจจุบันยังคงพบว่ามีผู้ที่ถูกโจมตีด้วย Phishing เป็นจำนวนมากขึ้นทุกปี ซึ่งสร้างความเสียหายให้แก่ผู้ตกเป็นเหยื่อเป็นอย่างมาก ในงานวิจัยเล่มนี้ผู้วิจัยได้ทำการศึกษาเฉพาะเนื้อหาข้อความภายใน Phishing E-Mails เท่านั้นโดยใช้กระบวนการ Text Classification System ในการวิเคราะห์เนื้อหาข้อความภายใน Phishing E-Mail ตาม Features ทั้ง 8 Features ที่ได้กำหนดไว้ โดยเก็บรวบรวมข้อมูล Phishing E-Mails จาก www.419scam.org ผลลัพธ์ที่ได้จะเป็นกฎที่ช่วยในการตัดสินใจ และเมื่อนำมาวัดประสิทธิภาพความแม่นยำของตัวชี้วัดพบว่า ตัวชี้วัดที่มีความแม่นยำสูงที่สุดคือ Binary Term Occurrences ซึ่งมีค่าความแม่นยำกว่า 80% หลังจากนั้นจะนำกฎการตัดสินใจที่ได้ไปวิเคราะห์ต่อด้วยการใช้ Association Rules เพื่อวิเคราะห์หาความสัมพันธ์ของแต่ละ features ซึ่งจะนำไปสู่การสร้างกฎสุดท้ายที่ใช้ในการพิจารณา Phishing E-Mail ดังตัวอย่างกฎที่ได้เช่น ถ้าอีเมล์ปรากฏข้อความตาม Feature ข้อความที่มีการแจ้งว่าอีเมล์นี้ต้องเป็นความลับเท่านั้น, ข้อความที่มีลักษณะเน้นย้ำให้รีบดำเนินการทันที และข้อความที่มีการขอความช่วยเหลือ ถือว่าเป็น E-Mail Phishing เป็นต้น ซึ่งงานวิจัยชิ้นนี้จะช่วยในการวิเคราะห์ข้อความภายใน E-Mail ว่ามีความเสี่ยงที่จะเป็น E-Mail Phishing หรือไม่ และเป็นส่วนหนึ่งที่จะลดความเสี่ยงจากการถูกโจมตีโดย E-Mail Phishing อีกด้วยในอนาคตควรจะขยายขอบเขตงานวิจัยด้วยการนำองค์ประกอบอื่นๆภายใน E-Mail เช่น ความน่าเชื่อถือของ Domain, ประเภทไฟล์ที่แนบมาใน E-Mail มาพิจารณาเพิ่มเติม

63 หน้า

CONTENTS

CONTENTS (cont.)

CONTENTS (cont.)

CONTENTS (cont.)

LIST OF TABLES

LIST OF TABLES (cont.)

LIST OF FIGURES

LIST OF FIGURES (cont.)

CHAPTER I
INTRODUCTION

1.1 Background and Problem Statement

Currently, there is an increasing trend of using internet continuously. The convenience of connecting to internet has been increased. It can be said that nowadays internet is an important factor for daily life of people. Internet has been used for business, communication, education and financial transactions. These activities relate to internet connection. Even though internet has been popularly used for many years and it has been continuously developed, the internet security is still a main concern issue. It is due to the fact that a group of people so-called Hacker creates an unwanted programs to attack computer system. The computer attack can cause the computer to stop working or thieve users' important and personal information, etc. Although tools for detecting the hacking have been developed, there are new channels and formats for hackers to attack the computer users.

Presently, the number of cyber attacks has been increased and the degree of cyber attacks has been increasing rapidly [1]. The growth of attack is in an exponential way. With various types of attacks and the regular change in attack methods, the number of users who are attacked on internet has also been increased. The impact of cyber attacks is not only on internet users but also business entities. It can cause a huge damage to business.

Phishing is another type of cyber attacks. It distracts internet users. It attacks users by using website or E-mail to obtain important user information. The obtained user information is then used for their benefits such as fake document or piracy. Phishing has been used by attackers for many years. A large number of internet users have still been attacked by phishing and the trend of users who are attacked by phishing has not been declined [2]. There are two main types of phishing.

1. Website Phishing

The fake website has been created. It imitates the real website by using font, logo, website format etc. If a user does not notice about these features, the user will not know that the fake website is being used. When the user uses the fake website by entering username and password, the user will be redirected to the designated page. The warning page will come up with a short message that the username and password entered are incorrect. This is to make the user misunderstand that it is a mistake typing. However, username and password have been collected by the attacker completely and used without authority or for illegal proposes.

2. E-mail Phishing

E-mails are send from a reliable organization such as bank, credit card company, etc. In the E-mail, the message asks users to update their personal information such as first name and last name, address, telephone number, identity card number, etc. Or the message says something that attracts users to pay attention to this phishing email. For example, the message informs that the email user wins the biggest prize or there is a very good offer to the email user. If the email user believes in the phishing email message, personal information will be stolen for illegal activities or the user could lose money.

Both phishing types are to be thieve the personal user information for the illegal proposes. If users know and can notice the important point, they can detect from phishing. Damages from phishing can be reduced. The detection of two phishing types are different. For Website phishing, when using a website, the user should notice about URL and evaluate components of URL whether it is like phishing website. For Email phishing, components in email should be evaluated whether it is phishing email. For example, who is email sender, file type attached with the email, and email messages. For email message, the user can notice about grammar, greetings, and sender's details including contact number and email. Campaign messages shown in email can notify the users whether it is phishing email.

This research analyzes types of phishing email to determine what invitation characteristics are used to attract users and what makes the user to understand that it is reliable email. The analysis is based on the campaign message in email only. The campaign message can be divided into 8 aspects. This is to define the

format of campaign messages in phishing email. After analyzing the content of campaign messages, the text classification system is created to analyze the characteristic, format and word usage in the campaign message of phishing email.

1.2 Objectives

This research applies Text Classification System to analyze contents in phishing email by using the designed features. The objective is to analyze content of phishing email to verify campaign messages of phishing email and to determine words used in message. Then, analysis results are used for detecting the phishing email.

1.3 Scope of Work

Phishing email has been gathered from www.419scam.org [3] where collects many phishing email messages. The analysis focuses only on contents in phishing email. There are eight factors being used for the content analysis.

1) Messages which offer the exaggerated propositions.
2) Messages which notify the recipients that the e-mail is confidential.
3) Messages which rush the recipients to take an immediate action.
4) Messages which ask for help.
5) Messages which inform the recipients that they get rewards even they don't participate
6) Messages which show only the words "Link", "Here" instead of URL
7) Messages which convince the recipients to donate or pay fees.
8) Messages which inform the recipients that their accounts has been limited.

After analyzing phishing email based on features, Text Classification System is applied to further analyze the campaign message in phishing email. Final results can be used for evaluating email message.

1.4 Expected Result

Expected results are for general email users who can analyze email contents. Eight features of campaign messages are applied for the analysis. This is to reduce risks and number of victims from phishing email.

CHAPTER II
LITERATURE REVIEW

For this research, Modular Crafting Text Classification Model for Phishing Email Detection, the author study the theories and relating literature for this study as follows.

2.1 Related Theory

2.1.1 Email

Email [4] stands for Electronic-Mail. Email system is derived from sending message to recipient between computer /and computer. A sender can type message on the computer and send this message via internet to a recipient. It does not require the recipient to receive the message during sending message. In addition, it does not require the recipient's computer to be turned on for receiving message. The recipient can view sent messages at any time.

Email address is used to locate Email message sender and recipient.

Components of E-mail address include three sections as the following format username@domainname.com

Username is the sender's account name. It identifies the E-mail owner. It could be a general name or organization name.

@ (At Sign) identifies the E-mail domain.

Domain name is the reference of mail server being used

2.1.2 Cybercrime

Cybercrime [5] is any action relating to computer that focuses on attack the computer system and information. It results in damages to victims but benefits go to attackers.

Cyber criminal is a person who accesses personal information via computer technology without permission. Information has been changed, copied or damages by the person.

Cybercrime evolution

From the beginning to now, cybercrime has the following details [6].

1) Privacy and personal information

In 1960's (1960 – 1970), it was the beginning to realize about the importance of cybercrime. During these years, there were many countries using computer to store their data and connect personal information database for governmental purposes. During the early years, cybercrime had not been recognized but it focused on the offence of personal information or classified. General users were mainly interested in and requested their government to protect personal information and professional confidential such as medical information, banking information, and etc.

2) Economy crime

In this era, the computer-related crime was categorized as economy crime or "White Collar Crimes." It means that criminals were white collars and reliable. They were educated and had ability. Economy crime covered the offence of damaging reliability and stability of economy and financial institutes, and etc.

In 1970's (1970 -1979), the government found statistic about cybercrime that affected the national economy. After the internet had been initially used in 1960 and the internet had been developed continuously, it had been used by the government and large business sectors. The new cybercrime was related to computer technology and internet network. These damaged economy. Cybercrime had been divided into two groups – 1. The criminal who uses computer as a tool and 2. The offence that the computer system and information in the computer were targets of the criminal.

Most crimes in this era that had been paid attention by academic and lawyers include

- Deception – it is the computer manipulation.

Computer manipulation was initially related to financial institutes, banks or company finance. The crime was about changing payment information, company financial reporting, or etc.

Computer manipulation was initially related to financial institutes, banks or company finance. The crime was about changing payment information, company financial reporting, or etc In the mid of 1980's, the computer manipulation was related to automatic teller machine, payment card, and credit card. The crime had been rapidly grown and expanded. The crime type was to steal ATM card and randomly choose password for cash withdrawal. The ATM card was stolen and changed its password by computer before cashing or etc.

In the tail of 1980's, the computer manipulation had been expanded to business other than bank, in particular telecommunication services via the internet network.

In 1990's, telecommunication service providers were main targets for the computer manipulation.

- Computersabotage & Computererpressung

Most crimes were related to personal computer by releasing virus or worm to destroy systems or information. This crime had been expanded after the increasing of internet network usage. The first virus had been written in 1986, namely "Pakistani Brain." Computer sabotage is crime significantly affecting the overall economy system in this era because computer and network system had been an increasingly important roles for routines of individual and business. Other than the computer sabotage, Computer Erpressung was another type of cybercrime. It was threatening for cash via E-mail. It asked to follow step by step specified unless the system would be destroyed or encrypted. Payment had been asked for the decryption.

- Computer hacking

In the beginning, it was not targeting to change or destroy information illegally but it was to test the ability to hack the security system. Presently, technology has been advanced. Hacking has also been developed. Targets are not only to hack the computer system for testing ability but also to use the system without payment or permission.

- Computerspionage

Since computer has been the large database, it can store a large number and various data. The perpetrator, then, looks for channels to access such data. Information espionage is to hack the system and access to databases. Tools are used to

detect data during the data transferring. This crime is often occurred in business and its counterparts.

● Piracy, Copy, or Software piracy

In the beginning, the target of these offences was to specific software. Presently, it increasingly focuses on personal computer. Targets have been moved to basic software because of its high price of copyright. The perpetrator does not prefer to pay for software. Thus, copy and software piracy have been used.

3. The publish of Illegal contents and cybercrime

From 1990's to present, cybercrime have not been covered only the violation of personal information but also other crimes such as publishing illegal contents, including gambling, pornography, defamation content, violence, and other via the website.

4. Other crimes

Other than the current crimes aforementioned, the traditional crime is linked to more computers. From the statistic, it shows that for many cases of computer information modification, the perpetrator did not only target on assets but attack the victim's life such as stealing or modifying medical information from hospital. In addition, the advanced technology can facilitate communication for criminal

2.1.3 Factors for successful Phishing

Phishing attack comes with various strategies that affect decision making of people [7]. For example,

● Reliable organization or person

An existing research about factors for successful Phishing found that most people trust and cooperate with the request from reliable organizations. Most E-mail Phishing are scams that refer to reputable organizations or entities

● Time limit

Phishing is monitored according to the time condition. It pressures the victim to reply quickly. It is to reduce time for the victim to check the Phishing E-mail.

● Tone

Phishing uses polite and official communication. It uses persuasive language to convince the victim's decision making. Such language can be seen in E-mail for greetings, subject, persuasion, or etc.

● Important information

The Phishing attacker creates the target to be attacked and collects important information of the victim such as name, work position, contact channel, and etc. Such information has been used to contact the victim to increase reliability and opportunity for successful attack.

● Fear

Phishing uses tactic to trick and make the victim trust in Phishing via the victim's fear. This is vulnerable to be attacked. Fear has been created to the victim by threatening the victim such as to limit the account usage or to use current situation and event for the trick. For example, disaster event has been used for raising funds and donations.

2.1.4 Impacts of phishing

● Waste from recovery

Data or account recovery from Phishing take times. It also creates confusion to the victim. Some cases take hours, months, or years for the recovery.

● Suffering

Phishing attack creates suffer, stress, anxiety to the victim. It can destroy trust and confidence of using the internet.

● Financial loss

Phishing attack can affect direct losses such as goods and service values and indirect losses such as legal consulting fees. For example, there was Phishing that created $1,343.12 loss to a victim.

● Social impact

Phishing attack creates the victim's stressful about personal relation such as friends and family. When the victim is attacked, not only personal but also personal relation information can be accessed. Such information can be used for the attack or creating conditions or targets for the victim to follow.

● Business impact

If an organization is attacked by Phishing, it could be followed by many big troubles. When the perpetrator can access customer information, the organization will take a serious risk that can affect its reliability and reputation. This can create a large number of losses to business.

● Waste of production

In business, when data are recovered from Phishing, it affects the efficiency of employees in the organization. Not IT but general employees will take more times to review the reliability when using each function to detect Phishing. This can affect the capacity of producing goods and services.

2.1.5 Phishing

Phishing [8] is a fake website that is similar to the real website. It is to deceive a user to fill in personal information and access to the system. Thus, there is risk for that account information to be stolen. The Phishing link is normally attached with email or there is invitation message to email recipients. The degree of damage from phishing does not affect to the recipient immediately. The recipient can suspend personal information such as changing the username and password when realizing that personal information has been entered through the fake website. However, damages could be occurred immediately in case of the user enters the personal information to the fake website and perform the online financial transaction.

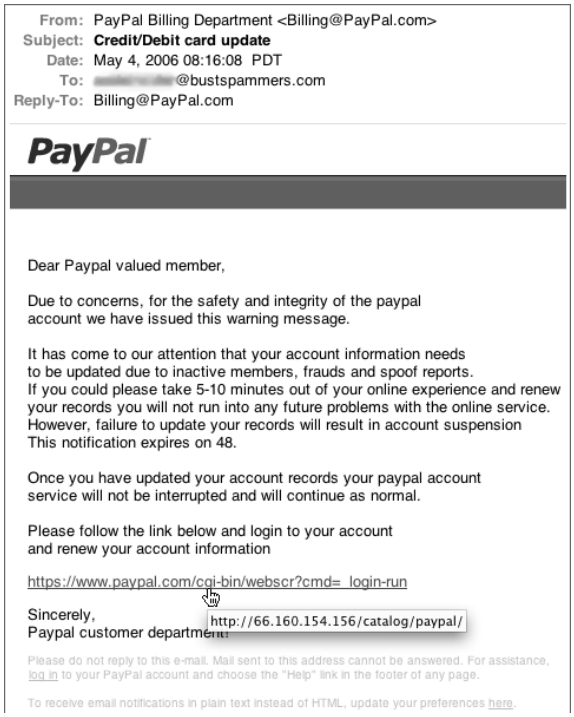
Phishing or Fishing means that the attack is like baiting. The attack is in from of the fake E-mail or website – Phishing E-mail or Phishing Website. The attack is occurred during important events such as disaster or natural disaster, that it has more chances for Phishing to be successful

Type of Phishing

1) Website Phishing is to create a fake website with components imitated from the real website. However, URL is created for information stolen. It is not the real URL. When a user enters username and password on the Phishing website, information will be stored and it leads to the theft of personal information

Figure 2.1 Website Phishing. [9]

2) E-mail Phishing is to create a fake E-mail. It makes the recipient to understand that the E-mail message has been sent from an organization. This is to deceive the email user to disclose personal information through the invitation or campaign message. The message makes the E-mail user think that it has been sent from a reliable sender


From: PayPal Billing Department <Billing@PayPal.com>
Subject: Credit/Debit card update
Date: May 4, 2006 08:16:08 PDT
To: @bustspammers.com
Reply-To: Billing@PayPal.com

PayPal

Dear Paypal valued member,

Due to concerns, for the safety and integrity of the paypal account we have issued this warning message.

It has come to our attention that your account information needs to be updated due to inactive members, frauds and spoof reports. If you could please take 5-10 minutes out of your online experience and renew your records you will not run into any future problems with the online service. However, failure to update your records will result in account suspension This notification expires on 48.

Once you have updated your account records your paypal account service will not be interrupted and will continue as normal.

Please follow the link below and login to your account and renew your account information

https://www.paypal.com/cgi-bin/webscr?cmd=_login-run

http://66.160.154.156/catalog/paypal/

Sincerely,
Paypal customer department

Please do not reply to this e-mail. Mail sent to this address cannot be answered. For assistance, log in to your PayPal account and choose the "Help" link in the footer of any page.

To receive email notifications in plain text instead of HTML, update your preferences here.

Figure 2.2 E-Mail Phishing. [10]

2.1.6 Classification Algorithm

It is a process of data mining used for classification. It classifies supervised information that objectives are clearly set. It is a prediction with a targeted solution. Information for classification has two sections as follows.

- Training set –it is to arrange data in a specified group. This stage obtains the classification model.
- Test set – After receiving the classification model, the test set is sent to the model to test and obtain results.

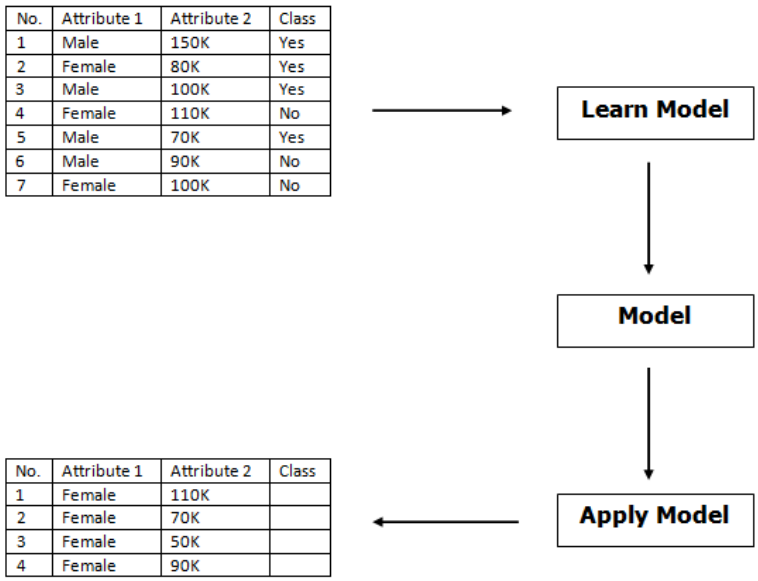


Figure 2.3 Classification Algorithm**.**

2.1.7 Text Mining

Text mining [11] is the mix of human and computer capacities. The human capacity is applied for language analysis. The computer capacity is applied for data mining. Text mining is a process of thinking analysis with different methods such as extracting or cutting words. This is to achieve new information

2.1.8 Tokenizing

Tokenizing [12] is the process of separating messages into words. It is to evaluate words in sentences and identify keywords.

2.1.9 Stemming

Stemming [13] is a process of deducing similar meaning words with different formats and present them in the same format. For example, presentation, presented, and presenting are streamed and they are deducted to be only “present.”

2.1.10 N-Grams

N-Gram [14] is a process of word sequence estimation in one sentence. N is 1 to n values. N-Gram is a popular model that it is an efficient model for language analysis

2.1.11 Stop Word Removal

For Stop Word Removal [15], there are many conjunctions in sentences. Those conjunctions deter Text mining process. They are not useful for word or document classification. Thus, Stop Word Removal can drop those conjunctions to reduce useless words. This can improve the efficiency of Text mining process.

2.1.12 TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) [16] is a process of weighting words. The weighting method is based on the frequency of using words in one document and all documents. The equation is a follows.

$$IDF = \log(\frac{N}{DF}) \quad (2.1)$$

2.1.13 Decision Tree

Decision Tree [17] is to classify data types into specific groups based on data characteristics. Data types are applied for decision making and analysis. Its component is Root Node that is the origin for decision tree. Each node has attribute value and branches of nodes are attached by values. Leave nodes present results from the data classification. The learning from decision tree helps us realize the importance of the Attributes. Decision tree helps us realize the importance of the attributes value. Users can take advantage of each attributes to analyze the data more efficiently**.**

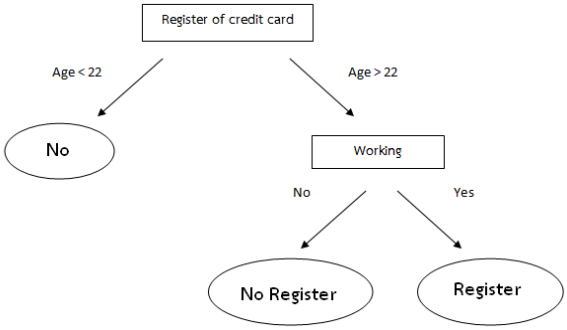


Figure 2.4 Decision Tree.

2.2 Related Literature

Danuwasin Charoen [18] studied Phishing: A Field Experiment from 174 MBA students whose accounts were registered to the university system. A phishing email imitated from the system officer's email was generated and sent to those students. The email content was about informing the students to check and modify their personal information. The phishing URL was attached with the phishing email. Results showed that 170 students accessing the phishing website and logging in the phishing website. Those student accounts who accessed the phishing website were then disclosed. Those students realized that the phishing email was reliable because the email was sent from the system administrator email that was included in the email list. This research analyzed that those students trusted the email list. Even though the attached URL was ambiguous, those students still collaborated to change their personal information. The research suggested that relating factors supporting the email reliability should be explored for detecting the phishing email.

Dhanalakshmi Ranganayakulu [19] studied about Detecting Malicious URLs in E-Mail – An Implementation. The URL attached with email was investigated. Other than URL analysis, suspected words often used in email were investigated. The research suggested that other components in email should be studied that those components include

- Verification of email sender list
- Verification of grammar accuracy in email
- Greetings used in email

- Messages written in email such as the exaggerated offer content or urgent message or personal information request, and etc.
- Email postscript
- Identify or contact details

Those components can be used to evaluate phishing email. This research focused on only the URL attached with phishing email. Most email did not present the URL but “Here” or “Click” for the recipient to click the link. After the click, the recipient then was redirected to the Phishing website.

CHAPTER III
RESEARCH METHODOLOGY

For this research, Content-Based Modular Crafting Text Classification Model for Phishing Email Detection, it aims to studying formats of campaign messages in phishing email. Those formats are analyzed to detect phishing email. Results from this research can be used to analyze the reliability of email and reduce a number of damages from phishing email. Thus, the author uses the following research methodology.

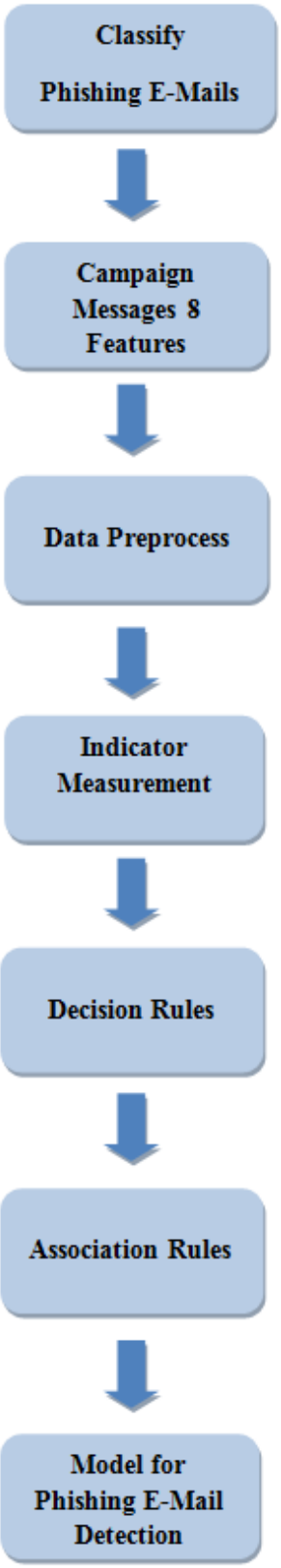


Figure 3.1 Diagram Process of Phishing Email Detection Model.

3.1 Business Understanding

Presently, there are a large number of continuing and increasing attacks. Those attacks are varied and they can cause damages from general users to business. One attack that have use is Phishing. This research focuses on the Email phishing by analyzing contents in email. Contents in phishing email include eight features. This is to analyze the format of contents in phishing email.

From the analysis of contents in phishing email, there are eight features as follows.

Table 3.1 Eight features of E-mail Phishing contents.

Features	Description	Example
C1 Messages which offer the exaggerated propositions	For this feature, the message is about to propose an email recipient including inviting to join business, asking for help with an exaggerated amount of compensation.	I am offering you 20% of the total funds for your assistance so show your willingness by replying to me through this email address (julianabrown59@yahoo.fr)
C2 Messages which notify the recipients that the e-mail is confidential	For this feature, the message is indicating that this message is confidential. It cannot be sent or informed to the other person. The reply message is required to be sent back to the specified email only.	We ask that you keep your Winning information confidential until your claims have been processed

Table 3.1 Eight features of E-mail Phishing contents.(Cont.)

Features	Description	Example
C3 Messages which rush the recipients to take an immediate action	This feature is related to other features such as the exaggerated offer message and winning prize message without joining activities. The message shows that an email recipient has already received or won the prize and it is required to contact back within a certain period of days or hours. This makes the recipient to feel that it is an urgent message. Then, the recipient does not analyze but believe in conditions created in the phishing email.	You have 48 hrs for you to send the requested fee.
C4 Messages which ask for help	This feature is about asking for help message. The sender is falling into various problems or severe sickness. The message has asked an email recipient for help and the email recipient will be compensated by different things. As a result, the campaign message makes the email recipient to believe that it is a true story and follow the steps set by the criminal.	I have a very desperate need for your help. I am seeking your kind assistance to move the sum of US$5.5 million

Table 3.1 Eight features of E-mail Phishing contents.(Cont.)

Features	Description	Example
C5 Messages which inform the recipients that they get rewards even they don't participate	This feature is about the message informing an email recipient who wins a big prize based on the random selection. The message mostly refers the big prize given from a large company.	Yahoo! Mail announces you as one of the 25 lucky winners in the ongoing 10 Years Yahoo lottery All 25 winning email addresses were randomly selected from a batch of 50,000,000 international emails Award of the New Year Held this month.
C6 Messages which show only the words "Link", "Here" instead of URL	This feature is a general message in email or there might be message contents as described previously. However, the email includes another feature that asking an email recipient to click the link rather than presenting URL. The email recipient should be careful to click Here. The email recipient does not know what URL that the email recipient will be redirected to. This feature can be related to another phishing that creates a fake website for deception.	If this email is not spam, click here to submit the signatures to FortiGuard

Table 3.1 Eight features of E-mail Phishing contents.(Cont.)

Features	Description	Example
C7 Messages which convince the recipients to donate or pay fees	This feature invites an email recipient to donate to different campaigns for helping people in disaster or inform that the email recipient wins a big prize but some fees applied before receiving the big prize.	IF YOU FAIL TO SEND THE $155 00 THIS WEEK YOUR $2.500,000.00 IS GONE
C8 Messages which inform the recipients that their accounts has been limited	This feature sends a message that a user's account has been limited. The user is then required to follow specified conditions to activate the account for normal uses. The specified conditions include the identification confirmation by asking for personal information, identification number, and etc. Personal information can create losses to the account owner.	It has come to our attention that your security information was recently changed. That requires you to verify your security information. Failure to validate your security information may result to account termination

3.2 Data Collection and Understanding

A source for phishing emails is taken from a website that collects different phishing emails. The website is www.419scam.org.

- Sample size

Samples are randomly selected for the analysis. The sample size is 500 Phishing E-mails.

- Data of each features are presented in chart

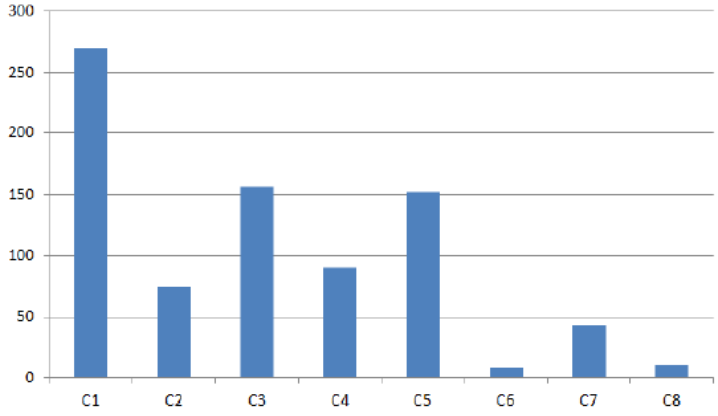


Figure 3.2 A number of emails in each feature found in 500 Phishing Emails.

3.3 Data Preprocessing

This process uses the analyzed email contents and separates them to each feature. Next, Tokenizing, Stemming, N-Grams, Stop word Removal and Term Frequency-inverse document frequency are performed to classify the contents in the similar format. Details of each process are described as the followings.

3.3.1 Tokenizing

This process cuts sentences and separates them to words as presented in Figure 3.2.

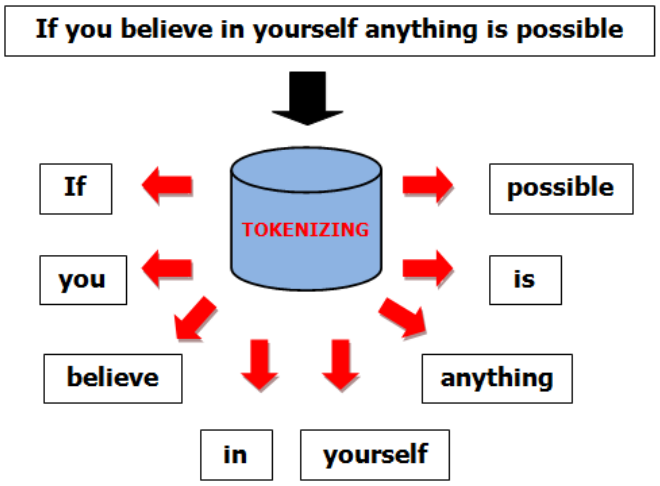


Figure 3.3 Tokenizing Process.

3.3.2 Stemming

This process is to search for root words. It is to analyze the words with similar meaning but their spelling is different. It also analyzes verbs, nouns, or adjectives that they are converted to root words. This process is presented in Figure 3.4

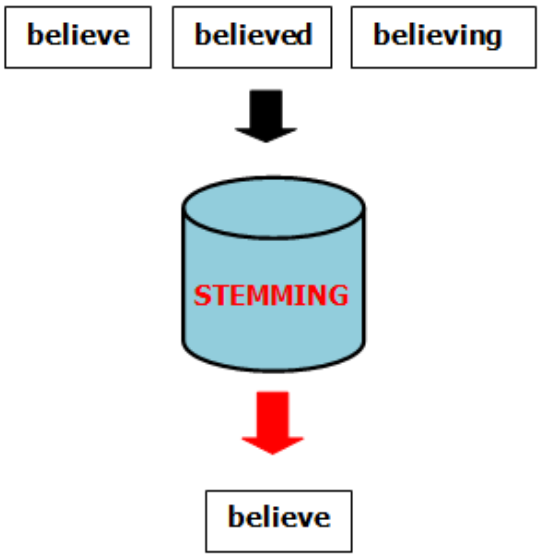


Figure 3.4 Stemming Process.

3.3.3 N-Grams

This process is to calculate the probability to create sentences or the connection of words. A sentence is separated into terms at N values. This research uses N equal to 2.

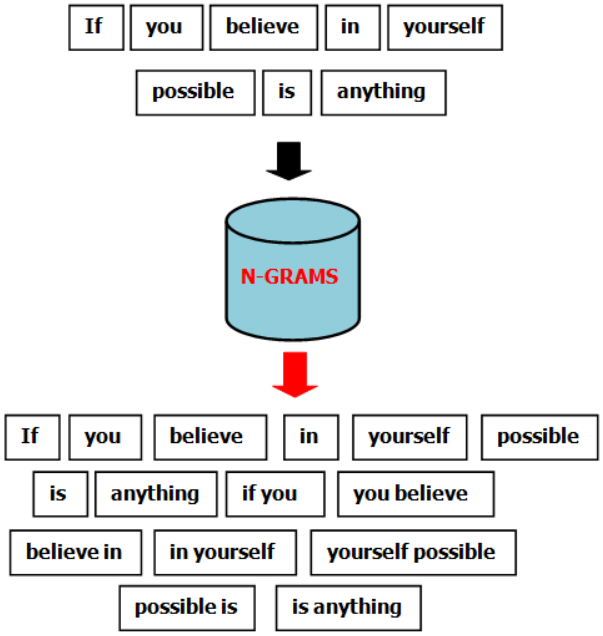


Figure 3.5 N-Grams.

3.3.4 Stop Words Removal

This process is to remove verbs, adjectives, and conjunctions such as is, am, are, to, and etc. This process is presented in Figure 3.6

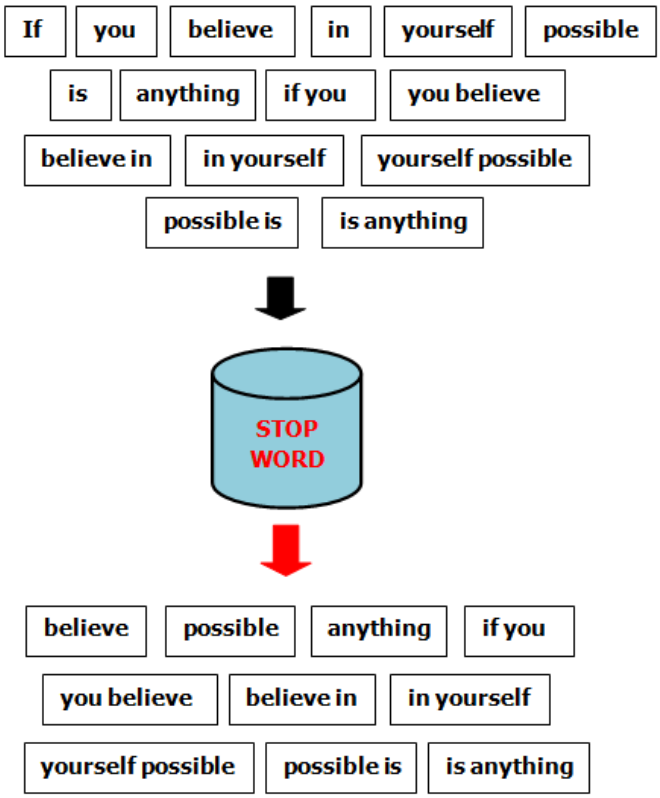


Figure 3.6 Stop word Process.

3.3.5 Indicator Measurement

This process is to compare the processed information. The accuracy of four measurements is as follows.

1. Binary Term Occurrences

This process is to weight by “including” or “not including” interested words in the document.

No.	Text
1.	I have urgent and very confidential business proposition for you
2.	Your help will be appreciated with 30% of the total sum
3.	We have made a random draw and your name and e-mail address was picked as the beneficiary to this Will. Please contact me urgently via my email
4.	The actual fees for shipping your ATM card is just $295
5.	I have a business proposal in the tune of $9.7m, (Nine Million Seven Hundred Thousand Dollars Only) after the successful transfer, we shall share in ratio of 40% for you and 60% for me. Please treat this business with utmost confidentiality and send me the Following
.	.....
500.	My offer to you is 10% of the total money after transfer and 50-50% share for both parties on any investment proceeds.

No.	confidential	total	address	urgent	offer	business	help
1.	1	0	0	1	0	1	0
2.	0	1	0	0	0	0	1
3.	0	0	1	1	0	0	0
4.	0	0	0	0	0	0	0
5.	1	0	0	0	0	1	0
.	...	...	...	...	...	...	...
500.	0	1	0	0	1	0	0

Figure 3.7 Binary Term Occurrences

From the above figure 3.7, based on 500 documents, it is 1 if the interested word is found in the document, otherwise 0

2. Term Occurrences

This process is to weight by frequency of interested words in the document.

No.	Text
1.	I have urgent and very confidential business proposition for you
2.	Your help will be appreciated with 30% of the total sum
3.	We have made a random draw and your name and e-mail address was picked as the beneficiary to this Will. Please contact me urgently via my email
4.	The actual fees for shipping your ATM card is just $295
5.	I have a business proposal in the tune of $9.7m, (Nine Million Seven Hundred Thousand Dollars Only) after the successful transfer, we shall share in ratio of 40% for you and 60% for me. Please treat this business with utmost confidentiality and send me the Following
.	.....
500.	My offer to you is 10% of the total money after transfer and 50-50% share for both parties on any investment proceeds.

No.	confidential	total	address	urgent	offer	business	help
1.	1	0	0	1	0	1	0
2.	0	1	0	0	0	0	1
3.	0	0	1	1	0	0	0
4.	0	0	0	0	0	0	0
5.	1	0	0	0	0	2	0
.	...	...	...	...	...	...	...
500.	0	1	0	0	1	0	0

Figure 3.8 Term Occurrences

From the above figure 3.8, based on 500 documents, it shows the number of words appeared in document.

3. Term Frequency

This process is to weight by frequency of interested words in the document and the total number of word.

No.	Text
1.	I have urgent and very confidential business proposition for you
2.	Your help will be appreciated with 30% of the total sum
3.	We have made a random draw and your name and e-mail address was picked as the beneficiary to this Will. Please contact me urgently via my email
4.	The actual fees for shipping your ATM card is just $295
5.	I have a business proposal in the tune of $9.7m, (Nine Million Seven Hundred Thousand Dollars Only) after the successful transfer, we shall share in ratio of 40% for you and 60% for me. Please treat this business with utmost confidentiality and send me the Following
.	.....
500.	My offer to you is 10% of the total money after transfer and 50-50% share for both parties on any investment proceeds.

No.	confidential	total	address	urgent	offer	business	help
1.	0.408	0	0	0.408	0	0.408	0
2.	0	0.378	0	0	0	0	0.378
3.	0	0	0.218	0.218	0	0	0
4.	0	0	0	0	0	0	0
5.	0.189	0	0	0	0	0.378	0
.	...	...	...	...	...	...	...
500.	0	0.316	0	0	0.316	0	0

Figure 3.9 Term Frequency

From the above figure 3.9 , based on 500 documents, it shows weights of each word compared to the total number of words in the document.

4. TF-IDF (Term Frequency-Inverse Document Frequency)

This process measures weights of words to determine the importance of Those words. It measures from the frequency of those words appeared in one document compared to the number of documents that have those words.

tf (Term frequency) is the frequency of words (t) appeared in document (d).

idf (inverse document frequency) is the inverse of frequency of document numbers that include those words. The following equation is applied.

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t} \tag{3.1}$$

N is a total number of documents

DF is a total number of documents that have the interested words (t).

TF-IDF is a weighted value for evaluating the importance of interested words appeared in a group of documents. The following equation is applied

$$tf - idf_{td} = tf_{td} \times idf_t \quad (3.2)$$

Example

From 500 documents,

idf of believe can be calculated by equation 3.1 as follows.

log (500/154) = 0.511

term		
believe	154	0.511
possible	92	0.735
anything	160	0.494
life	148	0.529
regret	254	0.294

tf-idf is calculated by equation 3.2. The below table shows the occurrence of words in each documents. Results are

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
believe	42	45	0	26	15
possible	5	0	35	0	24
anything	36	65	10	9	16
life	32	25	24	14	43
regret	55	24	16	46	23

Document No.1 : tf-idf of Believe = 0.511x 42 = 21.462

From five words explained, tf-idf is as follows.

believe = 21.462, 22.995, 0, 13.286, 7.665

possible = 3.675, 0, 25.725, 0, 17.64

anything = 17.784, 32.11, 4.94, 4.446, 7.904

if_you = 16.928, 13.225, 12.696, 7.406, 22.747

you_believe = 16.17, 7.056, 4.704, 13.524, 12.642, 6.762

After the comparison to measure the highest accuracy from four measurements, the highest accuracy measurement is selected to create the decision rule.

3.4 Campaign Messages

For the eight models of Classification Campaign Messages, Decision tree model is applied for classification to obtain decision rules. The rules are used for evaluating campaign messages in phishing email.

3.5 Evaluation

This process employs K-fold Cross-Validation method. This method tests the capacity of models to obtain the model reliability. This method divides data into different sections. This research specifies 10-fold cross validation. One fold is for test set for creating the set of models. The remaining folds are training sets for testing the model. Next, each dataset/fold is switched to be test set and training set for 10 times.

3.6 Phishing Message Modeling

This process will establish the rules of relation among features by using the Association Rules [20] to analyze the relation among features contained in an email. The rules of relation among features, resulted from the analysis, will help us to examine the e-mail content. If an e-mail contains more than one feature, it will be phishing. For example, if the email content is classified to the confidential message content, urgent message content asking for an immediate action, and request for help message content, these contents will be Phishing emails.

3.7 Research Schedule

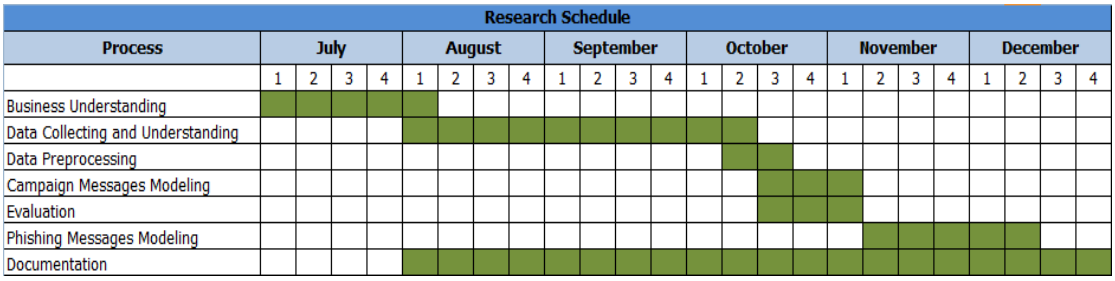


Figure 3.7 Research Schedule.

CHAPTER IV
RESULTS

4.1 Results from E-mail Phishing

From the sampling 500 E-mails for analyzing and classifying contents according to the 8 features. After the analysis, a number of data are classified as the followings.

Feature C1: Messages which offer the exaggerated propositions

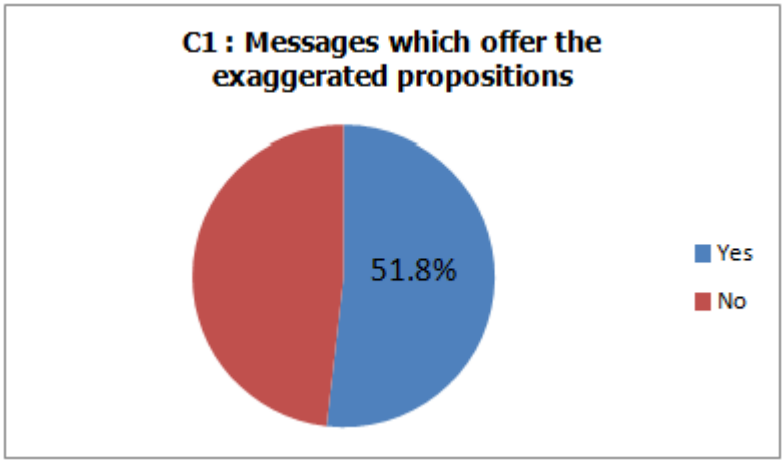


Figure 4.1 A number of messages that offer the exaggerated propositions.

Feature C2: Messages which notify the recipients that the e-mail is confidential

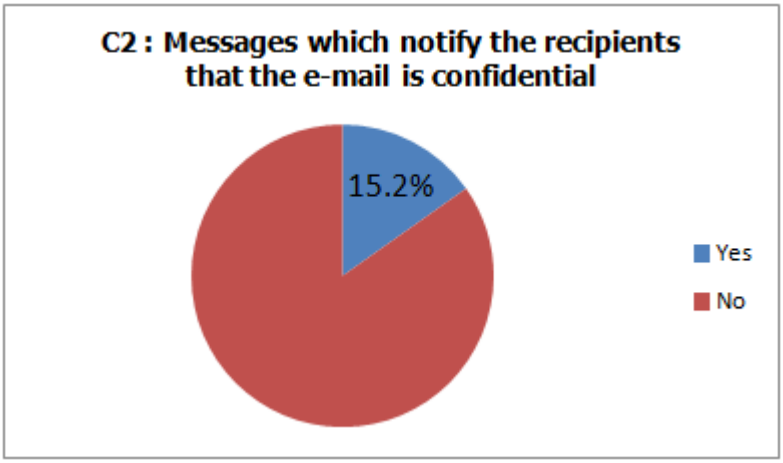


Figure 4.2 A number of messages which notify the recipients that the e-mail is confidential.

Feature C3: Messages which rush the recipients to take an immediate action

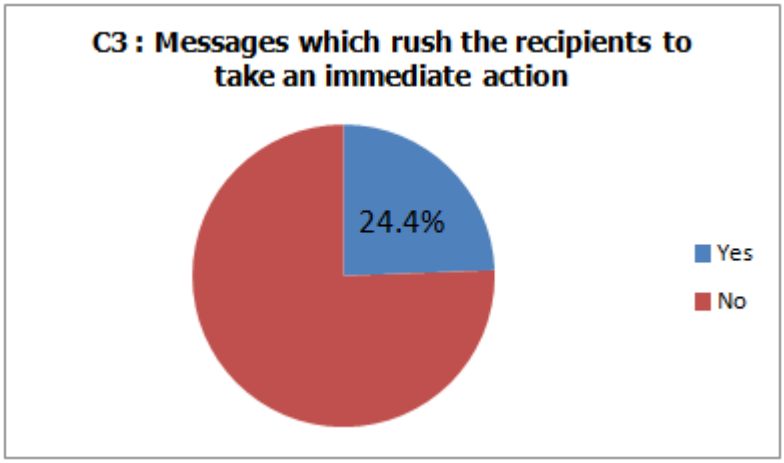


Figure 4.3 A number of messages which rush the recipients to take an immediate action.

Feature C4: Messages which ask for help

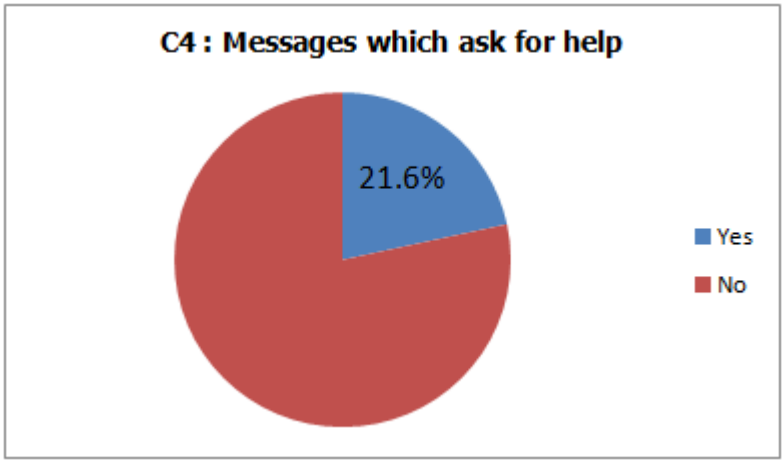


Figure 4.4 A number of messages which ask for help.

Feature C5: Messages which inform the recipients that they get rewards even they don't participate

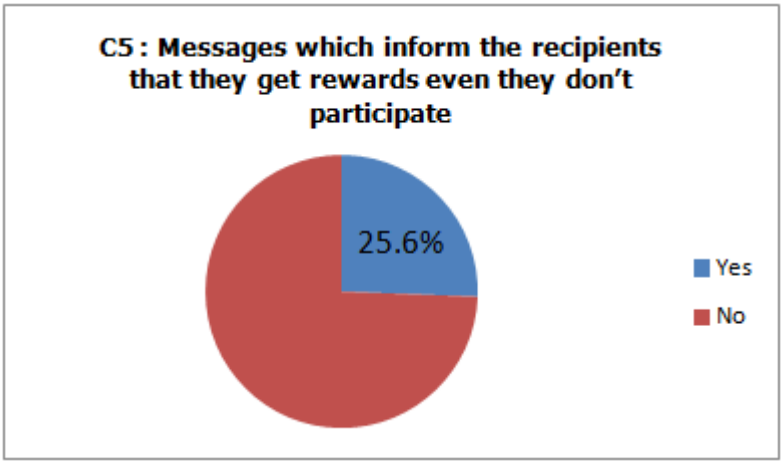


Figure 4.5 A number of messages which inform the recipients that they get rewards even they don't participate.

Feature C6: Messages which show only the words “Link”, “Here” instead of URL

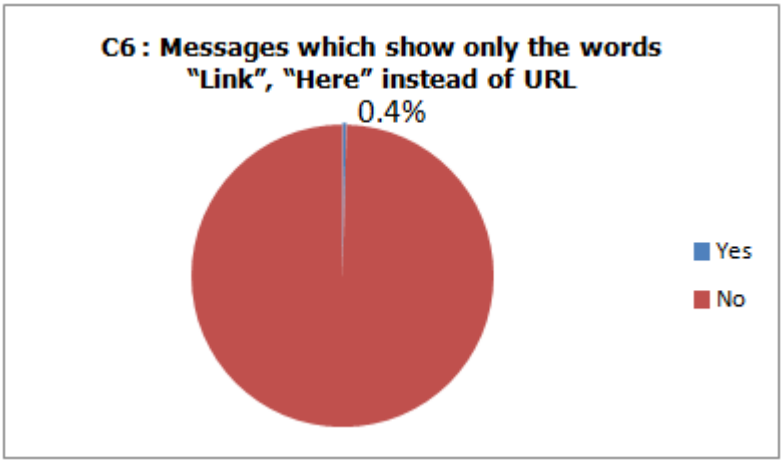


Figure 4.6 A number of messages which show only the word “Link”, “Here” instead of URL.

Feature C7: Messages which convince the recipients to donate or pay fees

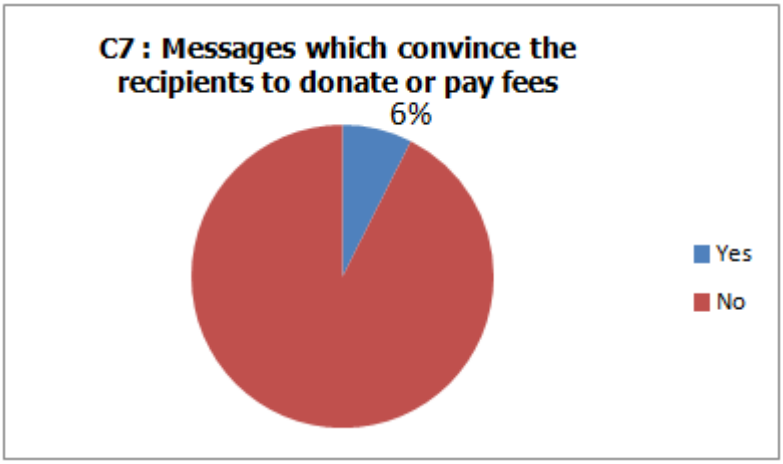


Figure 4.7 A number of messages which convince the recipients to donate or pay fees.

Feature C8: Messages which inform the recipients that their accounts has been limited

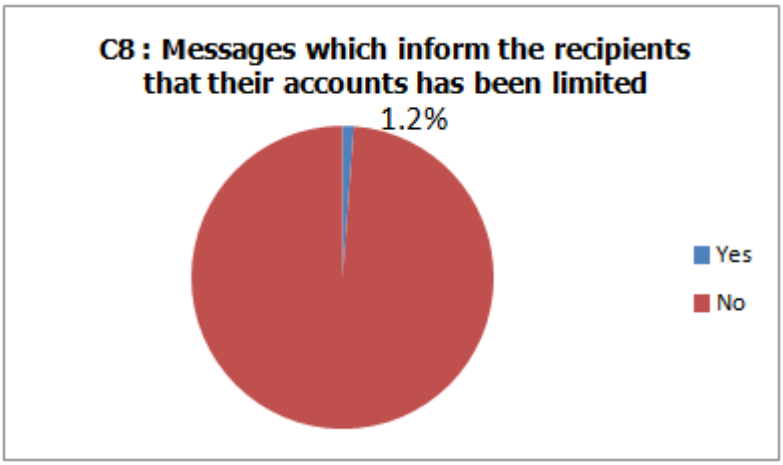


Figure 4.8 A number of Messages which inform the recipients that their accounts has been limited.

From 500 Phishing E-mails, the analysis shows that only six features can be processed through the decision tree analysis and rules for decision making will be created. From the above, it finds that contents of two features are 0 – 1.5 percent of 500 E-mails, including the Messages which show only the words “Link”, “Here” instead of URL and the Messages which inform the recipients that their accounts has been limited . Contents of both features are analyzed further rather than brought to the decision tree analysis. This is to further create rules for analyzing email phishing.

4.2 The Messages which offer the exaggerated propositions. (C1)

Contents in this feature inform about interesting offers or exaggerated returns to persuade an email user to agree with investing in business or providing supports.

4.2.1 Document measurement comparison

Table 4.1 Document measurement accuracy comparison of Feature C1.

	TF-IDF	Term Frequency	Term Occurrences	Binary Term Occurrences
Accuracy	67.40%	71.00%	70.60%	80.20%

From all four- measurement experiment, the result showed that the most accurate indicator is Binary Term Occurrences with 80.20%

4.2.2 Decision rules

```
won > 0.500: N
won ≤ 0.500
|  offer_you > 0.500: Y
|  offer_you ≤ 0.500
|  |  share > 0.500: Y
|  |  share ≤ 0.500
|  |  |  be_for > 0.500
|  |  |  |  kin > 0.500: N
|  |  |  |  kin ≤ 0.500: Y
|  |  |  be_for ≤ 0.500
|  |  |  |  while_will > 0.500: Y
|  |  |  |  while_will ≤ 0.500
|  |  |  |  |  expens > 0.500: Y
|  |  |  |  |  expens ≤ 0.500
|  |  |  |  |  |  total_amount > 0.500: Y
|  |  |  |  |  |  total_amount ≤ 0.500
|  |  |  |  |  |  |  you_while > 0.500: Y
|  |  |  |  |  |  |  you_while ≤ 0.500
|  |  |  |  |  |  |  |  distribut > 0.500: Y
|  |  |  |  |  |  |  |  distribut ≤ 0.500
|  |  |  |  |  |  |  |  |  will_like > 0.500: Y
|  |  |  |  |  |  |  |  |  will_like ≤ 0.500
|  |  |  |  |  |  |  |  |  |  ar_interest > 0.500
|  |  |  |  |  |  |  |  |  |  |  as_possibl > 0.500: N
|  |  |  |  |  |  |  |  |  |  |  as_possibl ≤ 0.500: Y
|  |  |  |  |  |  |  |  |  |  ar_interest ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  total_fund > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  total_fund ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  accept > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  accept ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  transfer_the > 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  and_i > 0.500: N
|  |  |  |  |  |  |  |  |  |  |  |  |  |  and_i ≤ 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  transfer_the ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  project > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  project ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  assist_me > 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  get > 0.500: N
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  get ≤ 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  assist_me ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  like_you > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  like_you ≤ 0.500: N
```

Figure 4.9 Decision Rules of Features C1.

Table 4.2 Description of Decision Rules of Feature C1.

No.	Included terms	Excluded terms
1.	“offer you”	“won”
2.	“share”	“won”, “offer you”
3.	“be for”	“won”, “offer you”, “share”, ”kin”
4.	“while will”	“won”, “offer you”, “share”, ”kin”, “be for”
5.	“expens”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”
6.	“total amount”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”
7.	“you while”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”
8.	“distibut”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”
9.	“will like”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”
10.	“ar interest”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”, “will like”, “as possible”
11.	“total fund”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”, “will like”, “as possible”, “ar interest”

Table 4.2 Description of Decision Rules of Feature C1. (Cont.)

No.	Included terms	Excluded terms
12.	“accept”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”, “will like”, “as possible”, “ar interest”, “total fund”
13.	“transfer the”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”, “will like”, “as possible”, “ar interest”, “total fund”, accept”, “and I”
14.	“project”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”, “will like”, “as possible”, “ar interest”, “total fund”, accept”, “and I”, “transfer the”
15.	“assist me”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”, “will like”, “as possible”, “ar interest”, “total fund”, accept”, “and I”, “transfer the”, “project”, “get”
16.	“like you”	“won”, “offer you”, “share”, ”kin”, “be for”, ”while will”, “expens”, “total amount”, “you while”, ”distribut”, “will like”, “as possible”, “ar interest”, “total fund”, accept”, “and I”, “transfer the”, “project”, “get”, “assist me”

4.3 Messages which notify the recipients that the e-mail is confidential. (C2)

For this feature, the message is indicating that this message is confidential. It cannot be sent or informed to the other person. The reply message is required to be sent back to the specified email only.

4.3.1 Document measurement comparison

Table 4.3 Document measurement accuracy comparison of Feature C2**.**

	TF-IDF	Term Frequency	Term Occurrences	Binary Term Occurrences
Accuracy	92.20%	92.20%	92.80%	93.00%

From all four- measurement experiment, the result showed that the most accurate indicator is Binary Term Occurrences with 93.00%

4.3.2 Decision rules

```
confidenti > 0.500: Y
confidenti ≤ 0.500
|   do_not > 0.500
|   |   assist > 0.500: Y
|   |   assist ≤ 0.500
|   |   |   fund_and > 0.500: Y
|   |   |   fund_and ≤ 0.500
|   |   |   |   with_the > 0.500: Y
|   |   |   |   with_the ≤ 0.500: N
|   do_not ≤ 0.500: N
```

Figure 4.10 Decision Rules of Features C2.

Table 4.4 Description of Decision Rules of Feature C2.

No.	Included terms	Excluded terms
1	“confidenti”	
2	“do not”, “assist”	“confidenti”
3	“do not”, “fund and”	“confidenti”, “assist”
4	“do not”, “with the”	“confidenti”, “assist”, “fund and”

4.4 Messages which rush the recipients to take an immediate action. (C3)

This feature is related to other features such as the exaggerated offer message and winning prize message without joining activities. The message shows that an email recipient has already received or won the prize and it is required to contact back within a certain period of days or hours. This makes the recipient to feel that it is an urgent message. Then, the recipient does not analyze but believe in conditions created in the phishing email.

4.4.1 Document measurement comparison

Table 4.5 Document measurement accuracy comparison of Feature C3**.**

	TF-IDF	Term Frequency	Term Occurrences	Binary Term Occurrences
Accuracy	76.20%	76.40%	79.00%	80.60%

From all four- measurement experiment, the result showed that the most accurate indicator is Binary Term Occurrences with 80.60%

4.4.2 Decision rules

```
soon_as > 0.500
|  total > 0.500: N
|  total ≤ 0.500
|  |  my_bank > 0.500: N
|  |  my_bank ≤ 0.500
|  |  |  and_your > 0.500: N
|  |  |  and_your ≤ 0.500
|  |  |  |  offer_you > 0.500: N
|  |  |  |  offer_you ≤ 0.500
|  |  |  |  |  share_the > 0.500: N
|  |  |  |  |  share_the ≤ 0.500
|  |  |  |  |  |  to_contact > 0.500: N
|  |  |  |  |  |  to_contact ≤ 0.500: Y
soon_as ≤ 0.500
|  urgent > 0.500
|  |  deal > 0.500: N
|  |  deal ≤ 0.500
|  |  |  will_give > 0.500: N
|  |  |  will_give ≤ 0.500
|  |  |  |  from_you > 0.500: N
|  |  |  |  from_you ≤ 0.500
|  |  |  |  |  that_will > 0.500: N
|  |  |  |  |  that_will ≤ 0.500
|  |  |  |  |  |  wish > 0.500: N
|  |  |  |  |  |  wish ≤ 0.500: Y
|  urgent ≤ 0.500
|  |  respond > 0.500
|  |  |  for_more > 0.500: N
|  |  |  for_more ≤ 0.500
|  |  |  |  want_to > 0.500: N
|  |  |  |  want_to ≤ 0.500
|  |  |  |  |  deal > 0.500: N
|  |  |  |  |  deal ≤ 0.500: Y
|  |  respond ≤ 0.500
|  |  |  respons > 0.500
|  |  |  |  confidenti > 0.500: N
|  |  |  |  confidenti ≤ 0.500
|  |  |  |  |  hundr > 0.500: N
|  |  |  |  |  hundr ≤ 0.500
|  |  |  |  |  |  ar_interest > 0.500: N
|  |  |  |  |  |  ar_interest ≤ 0.500
|  |  |  |  |  |  |  accept > 0.500: N
|  |  |  |  |  |  |  accept ≤ 0.500
|  |  |  |  |  |  |  |  organ > 0.500: N
|  |  |  |  |  |  |  |  organ ≤ 0.500: Y
|  |  |  respons ≤ 0.500
|  |  |  |  matter > 0.500
|  |  |  |  |  transfer > 0.500: N
|  |  |  |  |  transfer ≤ 0.500
|  |  |  |  |  |  becaus > 0.500: N
|  |  |  |  |  |  becaus ≤ 0.500: Y
```

Figure 4.11 Decision Rules of Features C3.

```
|  |  |  |  matter ≤ 0.500
|  |  |  |  |  as_possibl > 0.500
|  |  |  |  |  |  therefor > 0.500: N
|  |  |  |  |  |  therefor ≤ 0.500: Y
|  |  |  |  |  as_possibl ≤ 0.500
|  |  |  |  |  |  get_back > 0.500
|  |  |  |  |  |  |  your_email > 0.500: Y
|  |  |  |  |  |  |  your_email ≤ 0.500
|  |  |  |  |  |  |  |  am_mr > 0.500: Y
|  |  |  |  |  |  |  |  am_mr ≤ 0.500
|  |  |  |  |  |  |  |  |  of_million > 0.500: Y
|  |  |  |  |  |  |  |  |  of_million ≤ 0.500
|  |  |  |  |  |  |  |  |  |  offici > 0.500
|  |  |  |  |  |  |  |  |  |  |  address > 0.500: N
|  |  |  |  |  |  |  |  |  |  |  address ≤ 0.500: Y
|  |  |  |  |  |  |  |  |  |  offici ≤ 0.500: N
|  |  |  |  |  |  |  get_back ≤ 0.500
|  |  |  |  |  |  |  |  to_send > 0.500
|  |  |  |  |  |  |  |  |  to_the > 0.500: Y
|  |  |  |  |  |  |  |  |  to_the ≤ 0.500
|  |  |  |  |  |  |  |  |  |  claim > 0.500: Y
|  |  |  |  |  |  |  |  |  |  claim ≤ 0.500: N
|  |  |  |  |  |  |  |  to_send ≤ 0.500
|  |  |  |  |  |  |  |  |  it_to > 0.500
|  |  |  |  |  |  |  |  |  |  deposit > 0.500: N
|  |  |  |  |  |  |  |  |  |  deposit ≤ 0.500: Y
|  |  |  |  |  |  |  |  |  it_to ≤ 0.500
|  |  |  |  |  |  |  |  |  |  million_unit > 0.500
|  |  |  |  |  |  |  |  |  |  |  you_will > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  you_will ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  befor > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  befor ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  million_unit ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  that_we > 0.500
|  |  |  |  |  |  |  |  |  |  |  |  immedi > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  immedi ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  that_we ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  note > 0.500
```

Figure 4.11 Decision Rules of Features C3. (Cont.)

```
|  |  |  |  |  |  |  |  |  |  |  |  confidenti > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  confidenti ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  as_i > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  as_i ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  |  note ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  deliveri > 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  befor > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  befor ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  |  |  deliveri ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  have_to > 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  e > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  e ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  of_thi > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  of_thi ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  |  |  |  have_to ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  safe > 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  help_me > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  help_me ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  time ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  to_contact > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  to_contact ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  lucki ≤ 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  deposit_in > 0.500
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  chariti > 0.500: Y
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  chariti ≤ 0.500: N
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  deposit_in ≤ 0.500: N
```

Figure 4.11 Decision Rules of Features C3. (Cont.)

Table 4.6 Description of Decision Rules of Feature C3.

No.	Included terms	Excluded terms
1	“soon as”	“total”, ”my bank”, “and your”, “offer you”, “share the”, “to contact”
2	“urgent”	“soon as”, ”deal”, “will give”, “for you”, “that will”, “wish”
3	“respond”	“soon as”, ”urgent”, “for more”, “want to”, “deal”
4	“respons”	“soon as”, “urgent”, “respond”, “confidenti”, “hundr”, “ar_interest”, “accept”, “organ”
5	“matter”	“soon as”, “urgent”, “respond”, “respons”, “transfer”, “becaus”

Table 4.6 Description of Decision Rules of Feature C3. (Cont.)

No.	Included terms	Excluded terms
6	“as possibl”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “therefor”
7	“get back”, “ your email”	“soon as”, “urgent”, “respond”, “respons”, “matter”, and “as possibl”
8	“get back”, “ am mr”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “your email”
9	“get back”, “of million”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “ your email”, “am mr”
10	“get back”, “offici”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “ your email”, “address”
11	“to send”, “to the”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”
12	“to send”, “claim”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to the”
13	“it to”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “deposit”
14	“million unit”, “you will”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”
15	“million unit”, “befor”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “you will”
16	“that we”, “immedi”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”

Table 4.6 Description of Decision Rules of Feature C3. (Cont.)

No.	Included terms	Excluded terms
17	“note”, “confidenti”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, it to”, “million unit”, “that we”
18	“note”, “as i”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “confidenti”
19	“deliveri”, “befor”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”
20	“have to” and “e”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”, “deliveri”
21	“have to”, “of thi”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”, “deliveri”, “e”
22	“safe”, “help me”	soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”, “deliveri”, “have to”
23	“pai”, “pleas”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”, “deliveri”, “have to”, “safe”

Table 4.6 Description of Decision Rules of Feature C3. (Cont.)

No.	Included terms	Excluded terms
24	“lucki”, “time”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”, “deliveri”, “have to”, “safe” and “pai”
25	“lucki”, “to contact”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”, “deliveri”, “have to”, “safe”, “pai”, “time”
26	“deposit in”, “chariti”	“soon as”, “urgent”, “respond”, “respons”, “matter”, “as possibl”, “get back”, “to send”, “it to”, “million unit”, “that we”, “note”, “deliveri”, “have to”, “safe”, “pai”, “lucki”

4.5 Message which asked for help (C4)

This feature is about asking for help message. The sender uses a story that is falling into various problems such as severe sickness closing to death or disaster. The message has asked an email recipient for help and the email recipient will be compensated by different things. As a result, the campaign message makes the email recipient to believe that it is a true story and follow the steps set by the criminal such as transferring money for support or providing personal information.

4.5.1 Document measurement comparison

Table 4.7 Document measurement accuracy comparison of Feature C4**.**

	TF-IDF	Term Frequency	Term Occurrences	Binary Term Occurrences
Accuracy	79.60%	79.20%	78.00%	80.60%

From all four- measurement experiment, the result showed that the most accurate indicator is Binary Term Occurrences with 80.60%

4.5.2 Decision rules

```
your_assist > 0.500
|  you_in > 0.500: N
|  you_in ≤ 0.500
|  |  of_a > 0.500: N
|  |  of_a ≤ 0.500: Y
your_assist ≤ 0.500
|  help > 0.500
|  |  me_and > 0.500: N
|  |  me_and ≤ 0.500
|  |  |  and_you > 0.500: N
|  |  |  and_you ≤ 0.500
|  |  |  |  befor > 0.500: N
|  |  |  |  befor ≤ 0.500
|  |  |  |  |  immedi > 0.500: N
|  |  |  |  |  immedi ≤ 0.500: Y
|  help ≤ 0.500
|  |  safe > 0.500
|  |  |  i_want > 0.500: Y
|  |  |  i_want ≤ 0.500: N
|  |  safe ≤ 0.500
|  |  |  assist_me > 0.500
|  |  |  |  will_be > 0.500: N
|  |  |  |  will_be ≤ 0.500
|  |  |  |  |  in_thi > 0.500: N
|  |  |  |  |  in_thi ≤ 0.500: Y
|  |  |  assist_me ≤ 0.500
|  |  |  |  assist_to > 0.500
|  |  |  |  |  bank > 0.500: Y
|  |  |  |  |  bank ≤ 0.500: N
|  |  |  assist_to ≤ 0.500: N
```

Figure 4.12 Decision Rules of Features C4.

Table 4.8 Description of Decision Rules of Feature C4.

No.	Included terms	Excluded terms
1	“your assist”	“you are”, “of a”
2	“help”	“your assist”, “me and”, “and you”, “befor”, “immedi”
3	“safe”, “I want”	“your assist”, “help”
4	“assist me”	“your assist”, “help”, “safe”, “will be”, “in thi”
5	“assist to”, “bank”	“you assist”, “help”, “safe”, “assist me”

4.6 Messages which inform the recipients that they get rewards even they don’t participate.(C5)

This feature is about the message informing an email recipient who wins a big prize based on the random selection. The message mostly refers the big prize given from a large company.

4.6.1 Document measurement comparison

Table 4.9 Document measurement accuracy comparison of Feature C5.

	TF-IDF	Term Frequency	Term Occurrences	Binary Term Occurrences
Accuracy	82.00%	81.20%	83.40%	82.40%

From all four- measurement experiment, the result showed that the most accurate indicator is Term Occurrences with 83.40%

4.6.2 Decision rules

```
won > 0.500: Y
won ≤ 0.500
|   prize > 0.500: Y
|   prize ≤ 0.500
|   |   draw > 0.500: Y
|   |   draw ≤ 0.500
|   |   |   email > 2.500: Y
|   |   |   email ≤ 2.500
|   |   |   |   winner > 0.500: Y
|   |   |   |   |   bank > 4.500: Y
|   |   |   |   |   bank ≤ 4.500
|   |   |   |   |   |   with_the > 1.500: Y
|   |   |   |   |   |   with_the ≤ 1.500
|   |   |   |   |   |   |   you_ar > 2.500: Y
|   |   |   |   |   |   |   you_ar ≤ 2.500: N
```

Figure 4.13 Decision Rules of Features C5.

Table 4.10 Description of Decision Rules of Feature C5.

No.	Included terms	Excluded terms
1	“won”	
2	“prize”	“won”
3	“draw”	“won”, “prize”
4	“email”	“won”, “prize”, “draw”
5	“winner”	“won”, “prize”, “draw”, “email”
6	“bank”	“won”, “prize”, “draw”, “email”, “winner”
7	“with the”	“won”, “prize”, “draw”, “email”, “winner”, “bank”
8	“you are”	“won”, “prize”, “draw”, “email”, “winner”, “bank”, “with the”

4.7 Messages which convince the recipients to donate or pay fees. (C7)

This feature invites an email recipient to donate to different campaigns for helping people in disaster or donating to foundations. Or the message informs that the email recipient wins a big prize but some fees applied before receiving the authorized big prize.

4.7.1 Document measurement comparison

Table 4.11 Document measurement accuracy comparison of Feature C7**.**

	TF-IDF	Term Frequency	Term Occurrences	Binary Term Occurrences
Accuracy	97.60%	97.80%	97.20%	98.60%

From all four- measurement experiment, the result showed that the most accurate indicator is Binary Term Occurrences with 98.60%

4.7.2 Decision rules

```
fee > 0.500
|   confidenti > 0.500: N
|   confidenti ≤ 0.500: Y
fee ≤ 0.500
|   receiv_your > 0.500
|   |   as_soon > 0.500: Y
|   |   as_soon ≤ 0.500: N
|   receiv_your ≤ 0.500: N
```

Figure 4.14 Decision Rules of Features C7.

Table 4.12 Description of Decision Rules of Feature C7.

No.	Included terms	Excluded terms
1	“fee”, “confidenti”	
2	“receiv your”, “as soon”	fee

4.8 Messages which show only the words “Link”, “Here” instead of URL (C6)

This feature is a general message in email or there might be message contents in other features that are potentially phishing emails as previously described. For this feature, it can be noticed that there is a URL link in the content that will direct the user to a webpage when the user clicks Link or Here. The email recipient should be careful to click Here or Link rather than showing URL. The email recipient does not know what URL that the email recipient will be redirected to. This feature can be related to the website phishing that is one type of deception. The website phishing is to create a fake website for deceiving users.

4.8.1 Contents in Phishing emails for this feature, Messages which show only the words “Link”, “Here” instead of URL.

Table 4.13 Contents in Phishing emails of Features C6.

No.	Text
1	If this email is not spam, click here to submit the signatures to FortiGuard. I await your urgent response as we are working towards completing your payment this last quarter, as soon as this information is received, be rest assured that full attention will be given to beneficiary towards immediate release of approved funds.
2	After the last annual calculations of your fiscal activity we have determined that you are eligible to receive a tax refund of $63.80. Please submit the tax refund request and allow us 6 - 9 days in order to process it. A refund can be delayed for a variety of reasons, for example submitting invalid records or applying after the deadline. To access the form for your tax refund, please click here

4.8.2 Decision Rule

Table 4.14 Decision Rule of C6

No.	Included terms	Excluded terms
1	“click”, “here”	

4.9 Messages which inform the recipients that their accounts has been limited.(C8)

This feature sends a message that a user’s account has been limited to use. The user is then required to follow specified conditions to activate the account for normal uses. For example, the user has been asked to send personal information for authentication or update current personal information. If the user follows the phishing email message, personal information can create losses to the account owner.

4.9.1 Contents in Phishing emails for this feature, Messages which inform the recipients that their accounts has been limited.

Table 4.15 Contents in Phishing emails of Features C8.

No.	Text
1	It has come to our attention that your security information was recently changed. That requires you to verify your security information. Failure to validate your security information may result to account termination
2	We have recently determined that different computers have logged into your account and multiple password failures were present before the login. As a result of this, your account has been limited. We value the safety of our web-mail user and for these reasons your account has been selected carefully from a huge list as a subject of review. To start using your web-mail account again, you are being required to re-confirm your personal login details by clicking REPLY to this notice and provide the below information's to show that you are the rightful owner of this web-mail account.

Table 4.15 Contents in Phishing emails of Features C8. (Cont.)

No.	Text
	Warning!!! Account owner that refuses to re-confirm his/her account after 48 hours of receiving this notice will lose his/her account permanently
3	Due to the recent upgrade in linkedin you have to upgrade your account to keep using linkedin or your account will be terminated
4	Due to the congestion in all Webmail account and removal of all unused Accounts, we would be shutting down all unused accounts, You will have to confirm your E-mail by filling out your Login Info below after clicking the reply button, or your account will be suspended within 48 hours for security reasons.
5	I have waited enough to send your Cheque bank draft $2.8m united state US dollars for your compensation but I did not hear from you since then and the time the bank draft will be expired is getting near. Note: That you have to pay them for their Security Keeping Fee which is 185 Euro only.

4.9.2 Decision Rule

Table 4.16 Decision Rule of C8

No.	Included terms	Excluded terms
1	“account”, “terminate”	
2	“account”, “expire”	
3	“account”, “suspend”	

4.10 Rules for decision making

From all Features C1-C8 verified by four – measurement indicators, the most accurate values of each feature are as follows.

Table 4.17 The accuracy of Features C1-C8.

Features	Accuracy
C1	80.20%
C2	93.00%
C3	80.60%
C4	80.60%
C5	83.40%
C6	-
C7	98.60%
C8	-

From all Features C1-C8 set the rules of relation among features by analyzing with the Association Rules. The rules of relation among features are as follows.

Table 4.18 Rules for decision making and reliability.

No.	Rules	Confidence
1.	C2, C3, C4	1.00
2.	C2, C4	0.69
3.	C2, C3	0.68
4.	C1, C2	0.64
5.	C1, C4	0.57
6.	C3, C4	0.47
7.	C1, C3	0.39
8.	C3, C5, C7	0.38

Table 4.19 Rules for decision making.

No.	Rules	Description
1	C2, C3, C4	The confidential message, the urgent message for an immediate action, and the help request message are features that are defined as E-Mail Phishing.
2	C2, C4	The confidential message and the help request message are features that are defined as E-Mail Phishing.
3	C2, C3	The confidential message and the urgent message for an immediate action are features that are defined as E-Mail Phishing.
4	C1, C2	The message of exaggerated offer and the confidential message are features that are defined as E-Mail Phishing.
5	C1, C4	The message of exaggerated offer and the help request message are features that are defined as E-Mail Phishing.
6	C3, C4	The urgent message for an immediate action and the help request message are features that are defined as E-Mail Phishing.
7	C1, C3	The message of exaggerated offer and the urgent message for an immediate action are features that are defined as E-Mail Phishing.
8	C3, C5, C7	The urgent message for an immediate action, the winning prize message without joining the activity and convince the recipients to donate or pay fees message are features that are defined as E-Mail Phishing.

From this research, it examines 500 E-mails to analyze email contents. Those contents are classified to eight features. There are six features, including C1, C2, C3, C4, C5 and C7 that can be measured by four measurement tools. From the efficiency measurement of four measurement tools, it finds that Binary Term Occurrences provide the most accurate result because this measurement tool is less complexity relative to other methods. In addition, data used for the analysis are not

large. The other two features, including C6 and C8 analyze email contents for creating E-mail phishing rules rather than examine those contents with the measurement tools. After obtaining rules from eight features, rules are applied to determine the relationship between features. One email can be grouped into more than one feature. Then, the relationship rule is obtained. For example, an email content falling in Feature C2, C3 and C4 is considered as Phishing email or an email content falling in Feature C2 and C4 is considered as Phishing email.

CHAPTER V
SUMMARY AND RECOMMENDATION

5.1 Summary

Presently, the use of internet has been increasing continuously and rapidly. Internet has been used in different activities, including finance, business, communication, and etc. The more users are closely to internet, the more users are closely to attack. Website, file downloading, or emails are channels for the attack. The author studies different forms of attacks that can easily get through general users. Thus, this research is interested in Phishing attack. Phishing attack has been existed for many years and currently there are many users being attacked by Phishing. The author studies Email phishing and finds that email contents can be analyzed to determine the opportunity of that email to be phishing email. Thus, the detection of phishing email is proposed by analyzing email contents.

This research is about email phishing by analyzing contents of 500 Phishing emails. There are eight features for the email content analysis, including

C1. Messages which offer the exaggerated propositions.

C2. Messages which notify the recipients that the e-mail is confidential.

C3. Messages which rush the recipients to take an immediate action.

C4. Messages which ask for help.

C5. The Messages which inform the recipients that they get rewards even they don’t participate

C6. Messages which show only the words “Link”, “Here” instead of URL.

C7. Messages which convince the recipients to donate or pay fees.

C8. Messages which inform the recipients that their accounts has been limited.

. After analyzing email contents based on eight features, results of each features are further measured by four measurement tools, including TF-IDF, Term frequency, Term Occurrences and Binary Term Occurrences. Results can classify

contents into six features that have sufficient data to be further measured by four methods. Based on four measurement methods, Binary Term Occurrences provide the most accurate result at more than 80 percent. There are two features including C6 and C8 that have a low number of analyzed data. Then, they cannot be measured. However, the results from both features are further examined by the content analysis to determine the analysis rule of Phishing E-Mail. The relation among eight features is performed because one email can be classified into more than one feature. From the feature relation analysis, for example, if an email includes the Messages which notify the recipients that the e-mail is confidential (C2), Messages which rush the recipients to take an immediate action (C3), and Messages which ask for help (C4), the email is considered as Phishing E-Mail. If an email includes the Messages which notify the recipients that the e-mail is confidential (C2) and Messages which ask for help (C4), the email is considered as Phishing E-Mail. This research can help to analyzing the difference of email between reliable and phishing emails. Results can be used for analyzing phishing email contents that what feature has been used in email. It can be used to determine risk of email phishing and reduce risk of being attacked by the email phishing.

5.2 Limitation and recommendation

The limitation of this research includes 1) the Phishing E-Mail selected samples were manually analyzed their contents to classify those contents in 8 features. The analysis is based on the author. The classification is not based on any pattern but individual view. The classification method to each feature should be clearly identified and 2) it is about the number of data used in this research, a number of results for some features are low that those results cannot be further measured by the measurement tool and 3) there is no non-phishing email analyzed in this research. In the future, more samples should be applied and both phishing and non-phishing emails should be included in the analysis. Components in email other than the content should be analyzed such as the domain reliability, types of files attached to email, URL Link appeared in email, and etc. should be added as parts of analysis.

REFERENCES

1 ศูนย์ประสานการรักษาความมั่นคงปลอดภัยระบบคอมพิวเตอร์ประเทศไทย (ไทยเซิร์ต). (2559). สถิติภัยคุกคาม. Retrieved from https://www.thaicert.or.th/statistics/statistics.html

2 Anti-Phishing Working Group. (2016, Oct 03). APWG Phishing Attack Trends Reports. Retrieved from http://www.antiphishing.org/resources/apwg-reports/

3 joewein.de LLC. 2004 [cited 2016 Nov 24]. Available from:http//www.419scam.org

4 Warren Davies. What does an Email Address Mean? [cited 2016 Nov 24]. Available from: https://www.techwalla.com/articles/what-does-an-email-address-mean

5 Noparat Jitkornburi. 2010 [cited 2017 Jan 09]. Available From: http://oarnopparat.blogspot.com/2010/08/computer-crime.html

6 สาวตรี สุขศรี. 2009 [cited 2017 Jan 09]. Available From: https://facthai.wordpress.com/2009/11/09/

7 INA WANCA AND ASHLEY CANNON. [cites 2017 Jan 10]. Available From:http://www.nycrimecommission.org/pdfs/CCC-How-Human-Behavior-and-Decision-Making-Expose-Users-to-Phishing-Attacks.pdf

8 Mitesh Dedakia, Khushali Mistry. Phishing Detection using Content Based Associative Classification Data Mining. Journal of Engineering Computers & Applied Sciences (JECAS), Volume 4: No.7: July 2015.

9 *Phishing Quiz*[cited 2016 Nov 16]. Available from: https://www.opendns.com/phishing-quiz/

10 วิศัลย์ ประสงค์สุข. (27 เมษายน 2555). รู้จัก Phishing และการป้องกัน. Retrieved from https://www.thaicert.or.th/papers/general/2012/pa2012ge007.html

11 Vishal Gupta, Gurpreet S. Lehal. A Survey of Text Mining Techniques and Applications. Journal of emergint technologies in web intelligence, VOL. 1: NO. 1: AUGUST 2009

12 Dr.S.Kannan, Vairaprakash Gurusamy, Preprocessing Techniques for text Mining. October 2014

13 Ms. Anjali Ganesh Jivani. A Comparative Study of Stemming Algorithms. Anjali Ganesh Jivani et al, Int. J. Comp. Tech. Appl., Vol 2 : No.6: 1930-1938

14 Akarapol Ekwonganan. Identification of thai and transliterated words by n-gram models [Thesis for the Degree of Master of Arts in Linguistics]. Bangkok: Chulalongkorn University;2005

15 Murphy Choy. Effective Listings of Function Stop words for Twitter. International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 3: No. 6: 2012

16 Sutasinee Yokradubshan, Mahasak Ketcham. Application of Case-Based Reasoning an Online Helpdesk System for Website of Suan Sunandha Rajabhat University [abstract]. 2015

17 ชัดชัย แก้วตา, อัจฉรา มหาวีรวัฒน์. การวินิจฉัยคดีด้วยเทคนิคต้นไม้ตัดสินใจ[abstract].

18 ดนุวศิน เจริญ. Phishing : a field experiment. International Journal of Computer Science and Security (IJCSS), Volume (5): Issue (2): 2011

19 Dhanalakshmi Ranganayakulu. Detecting Malicious URLs in E-Mail – An Implementation[abstract]. 2013

20 กฤษณะ ไวยมัย. ธีระวัฒน์ พงษ์ศิริปรีดา, การใช้เทคนิค Association Rules Discovery เพื่อการจัดสรรกฎหมายในการพิจารณาคดีความ,NECTECT Technical Journal; Vol III;No.11

BIOGRAPHY

NAME	Miss Monthiya Sapan
DATE OF BIRTH	8 August 1991
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	University of the Thai Chamber of Commerce, 2009-2012 Bachelor of Science (Computer Science) Mahidol University, 2015-2017 Master of Science (Information Technology Management**)**
HOME ADDRESS	79/281 Soi Phetkasem 81/6 Nong-Khang-Phlu Nong-Khaem Bangkok 10160 Tel. 095-580-9666 Email: monthiya0808@gmail.com