

**DESIGN AND DEVELOPMENT OF MALAYSIAN CYBERSECURITY  
PROFILING FRAMEWORK: TOWARDS CREATING A RECOMMENDATION  
SYSTEM TO COMBAT CYBERCRIME**

**YEO HAN RONG**

**MASTER OF COMPUTER SCIENCE**

**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**UNIVERSITI TUNKU ABDUL RAHMAN**

**MARCH 2022**

## APPROVAL SHEET

This dissertation/thesis entitled “**DESIGN AND DEVELOPMENT OF MALAYSIAN CYBERSECURITY PROFILING FRAMEWORK: TOWARDS CREATING A RECOMMENDATION SYSTEM TO COMBAT CYBERCRIME**” was prepared by YEO HAN RONG and submitted as partial fulfillment of the requirements for the degree of Master of Computer Science at Universiti Tunku Abdul Rahman.

Approved by:



\_\_\_\_\_  
(Dr. AUN YICHJET)

Date: 01/03/2022

Supervisor

Department of Computer and Communication Technology

Faculty of Information and Communication Technology

Universiti Tunku Abdul Rahman



\_\_\_\_\_  
(Dr. JASMINA KHAW YEN MIN)

Date: 01/03/2022

Co-supervisor

Department of Computer Science

Faculty of Information and Communication Technology

Universiti Tunku Abdul Rahman



\_\_\_\_\_  
(Ts Dr. GAN MING LEE)

Date: 01/03/2022

Co-supervisor

Department of Computer and Communication Technology

Faculty of Information and Communication Technology

Universiti Tunku Abdul Rahman

**SUBMISSION SHEET**

**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**UNIVERSITI TUNKU ABDUL RAHMAN**

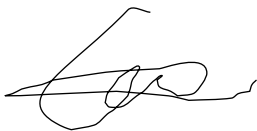
Date: 01/03/2022

**SUBMISSION OF DISSERTATION**

It is hereby certified that Yeo Han Rong (ID No: 19ACM05301 ) has completed this dissertation entitled “ **DESIGN AND DEVELOPMENT OF MALAYSIAN CYBERSECURITY PROFILING FRAMEWORK: TOWARDS CREATING A RECOMMENDATION SYSTEM TO COMBAT CYBERCRIME** ” under the supervision of Dr. Aun YiChiet from the Department of Computer and Communication Technology, Faculty of Information and Communication Technology, Dr. Jasmina Khaw Yen Min from the Department of Computer Science, Faculty of Information and Communication Technology, and Ts Dr. Gan Ming Lee from the Department of Computer and Communication Technology, Faculty of Information and Communication Technology.

I understand that the University will upload softcopy of my dissertation in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

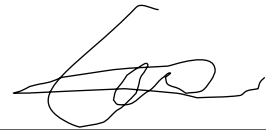
Yours truly,



(Yeo Han Rong)

## DECLARATION

I Yeo Han Rong hereby declare that the dissertation/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 UTAR or other institutions.



(Yeo Han Rong)

Date 01/03/2022

**DESIGN AND DEVELOPMENT OF MALAYSIAN CYBERSECURITY  
PROFILING FRAMEWORK: TOWARDS CREATING A RECOMMENDATION  
SYSTEM TO COMBAT CYBERCRIME**

By

**Yeo Han Rong**

A Dissertation submitted to the Department of Computer Science,

Faculty of Information and Communication Technology,

Universiti Tunku Abdul Rahman,

In partial fulfillment of the requirements for the degree of

Master of Science (Computer Science) in

August 2021

## **ABSTRACT**

### **DESIGN AND DEVELOPMENT OF MALAYSIAN CYBERSECURITY PROFILING FRAMEWORK: TOWARDS CREATING A RECOMMENDATION SYSTEM TO COMBAT CYBERCRIME**

**Yeo Han Rong**

Malaysia is deeply endangered by cybercrime and suffer huge economic losses. To against it, Malaysia government established institutions like MyCERT and CyberSAFE trying to increase the cyber security awareness of public by providing cyber security advice on the website but it not working as expected due to the ambiguity and complexity of the security advice. Moreover, evaluation of the cyber security awareness required a measurement method. Currently, most of the measurement method are using questionnaire and there is no measurement method that is applicable to an individual. This result in the awareness of public is unmeasurable. This research proposed a primary cybersecurity profiling framework for Malaysian by integrating Malaysia cybercrime situation, auto cyber security awareness measurement, and collaborative recommendation system. The users are first categorized into 3 levels based on their awareness result. The system will then provide recommendation to the users based on their awareness level. The profiling system is able to profile the users automatically using the data collected from users' computer. Using the matrix factorization collaborative filtering together with RNN-LSTM to train the model, the model with ISA-aware feature improved the RMSE by 0.12 and accuracy by 12.2%. Furthermore, the overall of cyber security awareness of users is improved after using the system.

## TABLE OF CONTENTS

	<b>Page</b>
APPROVAL SHEET	ii
SUBMISSION SHEET	iii
DECLARATION	iv
ABSTRACT	vi
LIST OF TABLES	x
LIST OF FIGURE	xii
CHAPTER	
1.0 INTRODUCTION	1
1.1 Overview	1
1.2 Problem Statements	3
1.3 Research Objectives	4
1.4 Contribution	5
1.5 Organization of thesis	6
2.0 LITERATURE REVIEW	6
2.1 Overview	6
2.2 Cybercrimes in Malaysia	7
2.3 Cyber Security Awareness	10
2.4 Security Awareness Measurement	11
2.5 Existing Security Awareness Measurement System	15
2.6 Recommendation System	18
2.6.1 Collaborative Filtering	18

2.6.2	Content-based Filtering	19
2.6.3	Knowledge-based	20
2.6.4	Hybrid Recommendation System	21
2.7	Recommendation	21
2.8	Summary	23
3.0	METHODOLOGY	23
3.1	Overview	23
3.2	Parameter composition	25
3.3	Security awareness measurement	28
3.4	Recommendation selection	30
3.5	Data collection	31
3.6	Modeling	33
3.7	Evaluation	37
3.8	Summary	38
4.0	RESULTS	38
4.1	Overview	38
4.2	Data collection	38
4.3	Training Results	42
4.4	Implications	47
4.5	Summary	48
5.0	CONCLUSION	48
5.1	Overview	48
5.2	Results	48
5.3	Contribution	48



5.4	Future Work	49
5.5	Summary	49
	REFERENCES	50

## LIST OF TABLES

Table 2.1: Type of cybercrime in Malaysia	8
Table 2.2: Sample of knowledge-based question in each focus area	12
Table 2.3: Question Type and Focus Areas of each measurement	14
Table 2.4: Features and Disadvantages of each existing work	15
Table 3.1: Focus Areas of modified HAIS-Q	26
Table 3.2: References and countermeasure of cybercrime	27
Table 3.3: Relationship between cybercrimes, focus area, and weight	28
Table 3.4: Equation variables definition	29
Table 3.5: Range of security awareness level	30
Table 3.6: Sub-area and checklist of password management	32
Table 3.7: Checklist of information handling	33
Table 3.8: Snippets of users-labelled dataset (rating)	34
Table 3.9: Unsorted user ratings for various security recommendations	36
Table 3.10: Ratings from end-users who are classified based on their ISA levels	36
Table 3.11: Confusion matrix	37
Table 4.1: Result of data collection	39
Table 4.2: Table of user security awareness level	41
Table 4.3: Average rating of each trigger recommendation	41
Table 4.4: Parameters of model training	42
Table 4.5: RMSE result of each batch size and learning rate with weight decay value 0.001	42
Table 4.6: RMSE result of each batch size and learning rate with weight decay value 0.01	43

Table 4.7: RMSE result of each batch size and learning rate with weight decay value 0.1	43
Table 4.8: RMSE result of each batch size and learning rate with weight decay value 1	44
Table 4.9: RMSE result of each batch size and learning rate with weight decay value 10	44
Table 4.10: RMSE result of model with and without features	45
Table 4.11: Confusion matrix of the model without features	46
Table 4.12: Confusion matrix of the model with features	46
Table 4.13: Confusion matrix of the model with feature on second profiling	47

## LIST OF FIGURE

Figure 1.1: The intuition for a holistic to cybersecurity solutions recommendations based on end user's IT literacy	3
Figure 1.2: Problem statement chart	4
Figure 3.1: System design overview	25
Figure 3.2: Flow diagram of data collection	31
Figure 3.3: Reinforcement learning using user ratings on security recommendations	34
Figure 3.4: 3-layers neural network	35

# CHAPTER 1

## INTRODUCTION

### 1.1 Overview

The number of cybercrime has exceeded ten thousand for four consecutive years since 2018. In 2018, there is RM 50 billion (USD 12.2 billion) of economic losses caused by the cybercrime which equal to 4% of Malaysia GDP, and the damage is keep increasing. Among the cybercrime, fraud is the vast majority in both cases and losses, followed by intrusion and malicious code. Fortunately, most on-site IT infrastructures are heavily secured through many layers of security like physical security, identity access management and networks security. However, the rise of work from home culture due to the current pandemic situation inevitably opened up new frontiers of security vulnerabilities. Some of the existing security policy is deemed infeasible as security moves from centrally coordinated implementations to free-roaming access. This recent shift of working paradigm ultimately pivoted some security decisions at the hands of users of different IT literacy. Users now have to connect remotely using personal devices on different networks that are not yet fortified. The lack of cyber-awareness and skillsets to operate recommended security solutions that range in usage complexity compounded this problem.

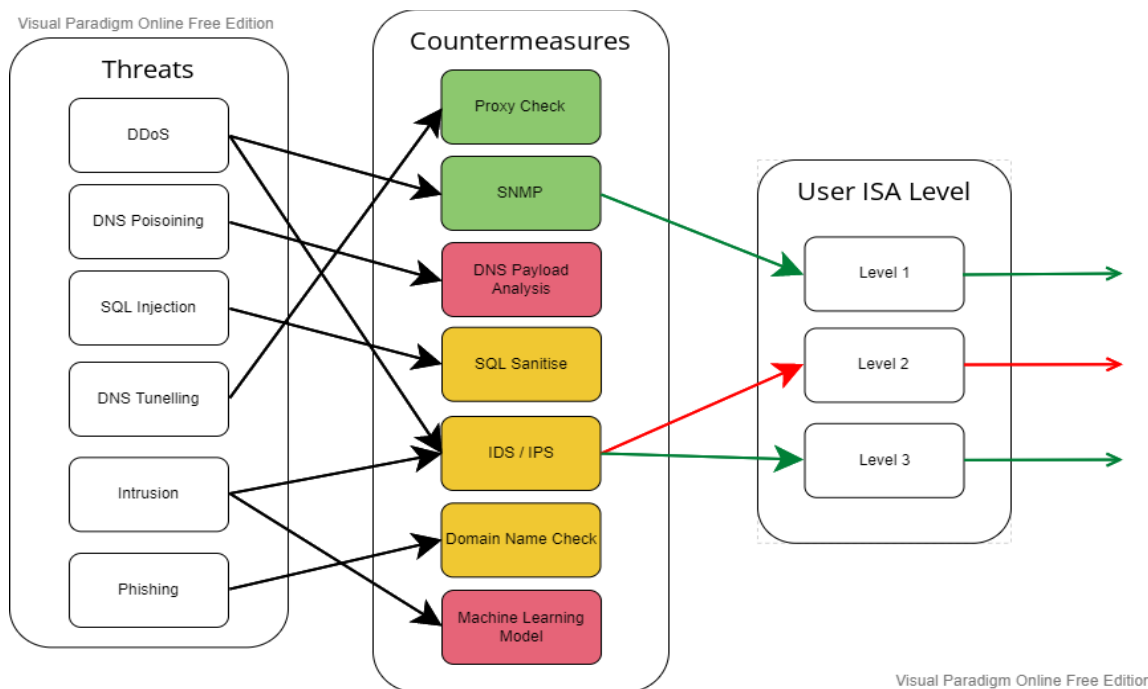
People could be the strongest defense against security threats with the right security awareness as most of the security vulnerabilities are caused by human factor (Kumaraguru et al., 2007). Therefore enhancing cyber security awareness can help to prevent cybercrime, as people know how to protect themselves from the cybercrime by applying to right security practices. Cyber security awareness is defined as the state of understanding of certain individuals against some potential cyber threats they are facing, and how to deal with them. In Malaysia, the Ministry of Communication and Multimedia Malaysia such as Malaysia Computer Emergency Response Team (MyCERT) and Cyber Security Awareness for Everyone (CyberSAFE) provided general cyber security advice to the public to improve security awareness and combat cybercrime. Unfortunately, most of these security warnings are complicated and ambiguous that dissuades people from applying security protection (Bada et al., 2019).

The researches that study cyber security awareness of different group of Malaysian (Ariffin & Letchumanan, 2020; Muniandy et al., 2017; Tan et al., 2020) gives the similar result that the cyber security awareness of Malaysian is at an average level and it is suggested to improve the awareness level to counter the threat of cybercrime, especially the phishing attack. Currently, cyber awareness can be measured using measurement methods like the Human Aspects of Information Security Questionnaire (HAIS-Q) or Information Security and Privacy Self-Assessment (ISPSA) (Galba et al., 2015; Parsons, McCormac, Butavicius, et al., 2014). The measurements are performed by distributing the questionnaire to the target audiences, and then calculates the score based on the answer of the target audience. The purpose is to determine the overall security awareness of an organization or the individuals in it. Based on the measurement result, the organization can act accordingly such as enhancing security policy or enhancing the security awareness of their employees. However, such the profiling method introduces the Hawthorn effect; where users react differently to the survey when the users are being monitored (Bada et al., 2019). Meanwhile, some of the security recommendations are event-based rather than user-based. This results in a gap in the ability of users needed to operate such tools in the optimal configurations for appropriate security advisories.

The constantly evolving traits of cyber threats made combating them a cat and mouse problem (Hart et al., 2020; Reep-van den Bergh & Junger, 2018; Von Solms & Van Niekerk, 2013). In network security, a range of countermeasures are designed to thwart different categories of network threats (Bendovschi, 2015; Sawaneh, 2020). For example, a network-wide visibility is needed for effective DDoS detection using SNMP; whereas SQL injection only needs selective hosts visibility. However, DDoS can be detected using only flow information while detecting SQL injection requires deep packet inspection, which means the user need the professional network knowledge to analysis the header and the content in the payload to check whether there is any SQL injection command. This paper hypothesizes that these security countermeasures are only as effective as the users who operate them. In addition to mapping solutions to threats; it is also important to map solutions to users based on their inherent technical proficiency. Figure 1.1 visualizes that security recommendations is an end-to-end process that starts with identifying the types of

threats; mapping effective countermeasures to thwart the threat and then finding fitting solutions to the end-users who are implementing the proposed solutions.

In many circumstances, users are prone to misconfigurations or even dissuaded from applying any security solutions due to unafforded operational complexity. For example, although IDS/IPS is commonly used to prevent network intrusions; the process required prior configurations like applying customized SNORT rules to be effective. As a result, this recommendation is less fitting for users who are less literate (see dotted red line, level2 users) than a cyber-savvy (green line, level 3 users). This holistic approach set user as the focal point instead as the weak-link for an end-to-end for security recommendations pipeline.



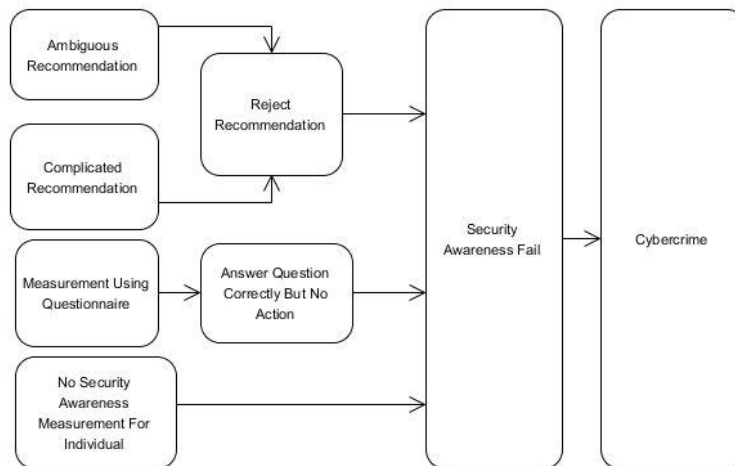
**Figure 0.1: The intuition for a holistic to cybersecurity solutions recommendations based on end user's IT literacy**

## 1.2 Problem Statements

Currently, the main problems to increase the ISA of public are the way our government provides the security advice is not effective, and there is no ISA measurement for public so that people do not know the security vulnerabilities of the majorities. Provide security advice by listing out all of them at once will have the opposite effect because it makes the

advice ambiguous. The advice is also not working while people think it is too complicated. The complexity of security advice can be subjective, and may relative to the security awareness level as the security awareness level reflects the ability of an individual in the cyber security domain. For the above reasons, the security advice provided on government’s website is no effective. Recommendation systems are able to solve this kind of problem by classifying users using user profiles, and predict what advice with the right complexity that the individual can accept. However, recommendation system requires user profile. It leads to another problem that is there is no ISA profiling system for public. Although there are some ISA measurement methods currently, most of the methods require questionnaire. Measuring the security awareness using a questionnaire may not accurate because people may not act accordingly even they choose the right answer in the questionnaire. In summary, there are 4 problems. The relationship between the problems and cybercrime is stated as shown in Figure 1.2.

- Ambiguous recommendations lead people to abandon security practices.
- Complicated recommendations lead people to abandon security practices.
- People do not take action even giving the right answer in the questionnaire.
- No security awareness measurement for the individual.



**Figure 0.2: Problem statement chart**

### 1.3 Research Objectives

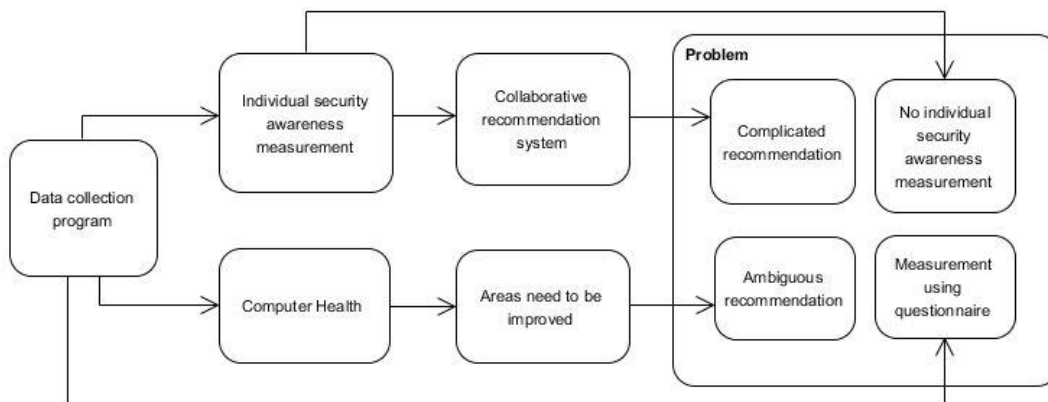
There are 2 main objectives for this research:



- Design a ISA profiling system for individual without questionnaire
- Develop a recommendation system that provide security advice based the security awareness level.

Firstly, the ISA profiling system collects the information from an individual’s computer as the measures of security awareness measurement. The user profile is generated using the data collected from user’s computer. The profile contains the necessary data to train the recommendation system.

Secondly, the recommendation system then classifies individuals into levels. The recommendation system uses the security awareness level to provide security advice to the individual. By giving the recommendation system a list of security advice, it is able to provide suitable security advice for certain types of people based on the user profile. The recommendation system is meant to increase an individual’s security awareness by providing the right security advice. The research objective chart shows how the objectives address the problems. (Figure 1.3)



**Figure 1.3: Research objective chart**

### 1.4 Contribution

Many measurements for security awareness have been created in the past. All of the previous measurements require the target audience to answer questions by using a questionnaire. Most of the measurements are toward measuring the security awareness of employees in an organization instead of the individual. No previous measurement taking the cases of cybercrime as measurement weight into account. This research proposed a

security awareness measurement that combines cybercrime in Malaysia and HAIS-Q to measures an individual's security awareness without using a questionnaire.

Recommendation systems have been used in the cyber defense domain as attack predictor and security advice provider. No previous recommendation system has combined security awareness as a user profile with a recommendation system to provide security advice. This research proposed a recommendation system that provides suitable security advice to users by classifying users using their security awareness.

## **1.5 Organization of thesis**

This thesis has 5 chapters in total: Chapter 1 introduction; Chapter 2 literature review; Chapter 3 methodology; Chapter 4 result and analysis; and Chapter 5 conclusion. Chapter 2 describes the literature review regarding the security awareness measurement and recommendation system. Chapter 3 explains the methodology of this research including parameter composition, security awareness measurement, recommendation selection, data collection, modeling, and evaluation. Chapter 4 analyzes the results of the recommendation system and examines the reliability. Finally, Chapter 5 summarizes the research and concludes for future study.

## **CHAPTER 2**

### **LITERATURE REVIEW**

#### **2.1 Overview**

To build a reliable cybersecurity profiling framework for Malaysian, the literatures that related to profiling system and recommendation system will be discussed in this chapter. The keywords searched including: cyber security awareness, information system awareness, cybersecurity awareness measurement, cybercrime Malaysia, cybercrime COVID-19 pandemic, recommendation system, collaborative filtering, content-based filtering, knowledge-based filtering, and hybrid recommendation system. The literatures are obtained from information technology related sources such as Institute of Electrical and Electronics Engineers (IEEE), ScienceDirect, Scopus, and ReserachGate.

The literatures that related to cybercrimes in Malaysia, security awareness, security awareness measurement, recommendation system, recommendation, and a summary will

be discussed. The types and impact of the most happened cybercrime in Malaysia will be explored. Definition and usage of security awareness will be introduced. The literature approaches of security awareness measurement and existing security awareness measurement systems will be reviewed.

Moreover, the types of recommendation systems which including collaborative, content-based, knowledge-based, and hybrid will be explored. The recommendation on the cybersecurity domain for the public will be discussed. Finally, the summary of this chapter is presented.

## **2.2 Cybercrimes in Malaysia**

With the rapid development of the Internet, the Internet is becoming more and more important in modern life. Due to its powerful data transmission function, a lot of activities can be done through Internet including E-commerce, banking, gaming, entertainment, and data transmission (Mui et al., 2002; Ruzgar, 2005). Unfortunately, following the dramatic growth of Internet usage, cybercrime is on the same trajectory, it growth rapidly too. Cybercrime causes damage to the economy and public security of a country, including Malaysia.

In 2002, University Technology Mara and the Malaysian Parliament were attacked. The hackers cleaned up all the information on the Parliament website and then replaced it with a foreign language. Then, a web defacement activities attacked many local websites in 2004. Malaysia has set up a series of laws to combat cybercrime such as the Computer Crimes Act 1997 (CCA), fining not exceeding RM150000 (USD35886) or to imprisonment for a term not exceeding ten years or to both; Communications and Multimedia Act 1998 (CMA), a fine, imprisonment up to one year, and additional fines for “every day or part of a day during which the offence is continued after conviction”. However, the cases of cybercrime are still increasing over years. In Malaysia, the cases of cybercrime had a 416% growth from 1139 cases in 2007 to 4738 cases in 2012, and it costs a total of RM286.2 million (USD 69 million) over 6 years from 2007 to 2012 (DSP Mahfuz Bin Dato’ Ab. Majid, 2013). Microsoft and Frost & Sullivan Study reveal that the potential economic losses caused by cybercrime can be as high as RM 50 billion (USD12.2 billion), which is more than 4% of Malaysia's GDP (Dashika Gnaneswaran, 2018).

Recently, due to COVID-19 pandemic, work from home culture rising. People have to work internet-based inevitably and it increases the chance for cybercrime. In fact, the threats of cybercrime actually increased during the pandemic across the world including Malaysia (Naidoo, 2020; Tharshini et al., 2021). The biggest cybercrime in Malaysia, fraud or said phishing, get even more dangerous during the pandemic as people switch the offline transaction method to online method such as bills and online shopping (Tharshini et al., 2021).

According to Malaysia Computer Emergency Response Team (MyCERT) official incident statistics from 2017 to 2019, the top 8 categories of cybercrime recorded in Malaysia are content-related, cyber harassment, denial-of-service, fraud, intrusion, intrusion attempt, malicious code, and spam. Table 2.1 shows the definition of each cybercrime defined by MyCERT.

**Table 2.1: Type of cybercrime in Malaysia**

<b>Cybercrime</b>	<b>Definition (by MyCERT)</b>
Content-Related	Any offensive, morally improper, and against current standards of accepted behavior including nudity and sex.
Cyber Harassment	A wide range of offensive behavior usually intended to disturb or upset and sometimes can be found threatening or disturbing.
Denial-of-Service	An attempt to make online service unavailable with traffic from multiple sources.
Fraud	Identity theft, stolen bank account, stolen passwords, stolen intellectual property and etc.

Intrusion	An unauthorized enter a computer, system, or network to access information and manipulate or render as system unreliable or unusable.
Intrusion Attempt	A potential unauthorized attempt to enter a computer, system, or network to access information and manipulate or render as system unreliable or unusable.
Malicious Code	Various forms of harmful software that intentionally designed to harm computer, network, and server.
Spam	Junk email.

Fraud, intrusion, and malicious codes are the three main cybercrime in Malaysia. According to the general incident statistics from MyCERT, the sum of fraud, intrusion, and malicious code is accounted for 83% of all cases in 2017, 95% in 2018, and 89% in 2019. Most of the cybercrimes that happened are categorized as fraud, intrusion, and malicious codes. Figure 2.1 shows the pie chart of cybercrime in Malaysia between the years 2017 to 2019.

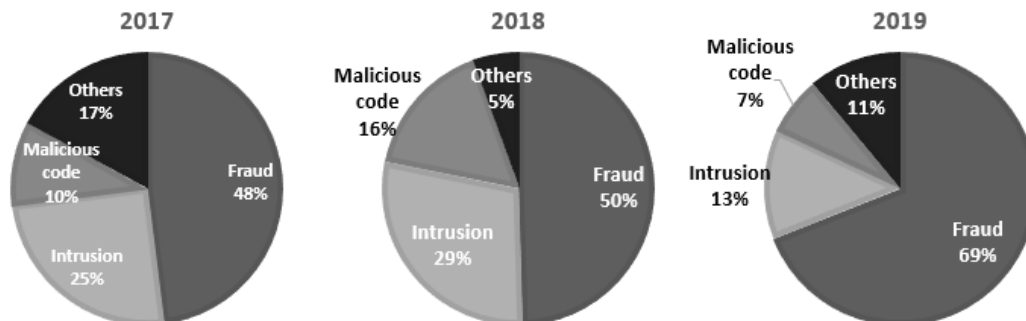


Figure 2.1: Pie chart of cybercrime in Malaysia between years 2017 to 2019

The statistics above show the distributions among the cybercrimes in 2017 to 2019 are biased. It may be because of the ease of some cybercrime to be implemented, the cyber

security weakness of Malaysian, or any other reasons. Anyway, the distribution of the cybercrime should be introduced into the calculation of ISA awareness as weight parameter because of the purpose of ISA measurement is to find out which security area needs to be enhanced.

In addition to using legislation to combat cybercrime, the government also set up originations and websites which aim to increase people's attention to security awareness such as CyberSAFE and MyCERT.

### **2.3 Cyber Security Awareness**

In daily life, there may be traps for cybercrime everywhere. People could easily fall into the trap inadvertently. It could be opening an attachment that is attached in an email, it could be plugging in a USB drive to the computer in a printing shop, and it also could be posting personal information on social media. Most of the time, the human factor is the main cause of cyber breaches (Kumaraguru et al., 2007; Shaw et al., 2009). Many cybercrimes can be prevented if people take seriously to security awareness.

The definition of cyber security awareness was defined by Shaw (Shaw et al., 2009) as: "Degree of understanding of users about the importance of information security and their responsibilities and acts to exercise sufficient levels of information security control to protect the organization's data and networks". Cyber security awareness does not only refer to the understanding or knowledge about cyber security, but also taking action of security protection. From the definition, the behaviors of individuals in the cybersecurity domain are able to reflect security awareness. So determining the security awareness by collecting the information from a computer that reflects an individual's behaviors is feasible.

There are many researchers study about the methods to increase cyber security awareness such as a security awareness training program, or a campaign. (Bada et al., 2019; Frey, 2018; Zwillig et al., 2020). The training program and campaign are welcomed by organizations and companies because it has been proved that enhancing security awareness is an effective method to prevent cybercrimes. The company send their employees to the training in order to increase security awareness, to protect the organization or company

from cybercrime. However, the security awareness training is more toward designed for organization and company but not for an individual.

Providing recommendations on the cybersecurity domain such as security advice and good security practices is a way to increase an individual's security awareness. However, the effect of recommendation may not be worked as expected because people reject the recommendation which they felt it is complicated or ambiguous (Bada et al., 2019). For example, a recommendation on firewall asks a person who does not know much about computer to configure the firewall at the port level in order to protect attack from the network. This will lead the person to lose the effort to apply the security practices.

## **2.4 Security Awareness Measurement**

Cyber security awareness measurement is a method that able to determine the security awareness of the target and convert the security awareness into numbers. Many papers study the methodology for information security awareness measurement. Since there are no universal standard for information security awareness measurement, researchers come out with different factors to determine security awareness. Typically, certain focus areas must be determined in order to perform security awareness measurements. The focus areas define the areas where the questionnaire should focus on to collect the necessary information.

Rahman, Lubis, and Ridho (Rahman et al., 2015) calculate the security awareness of a user by measures the following 11 areas, which are Self Attitude (SA), Self Behavior (BV), Self Cognitive (CT), Intention to Comply (IC), Policy Compliance (PC), Training Program (TP), Perceived Threats (RT), Inf. Security Awareness (IS), Peer Performance (PP), Social Pressure (SP), and Religious Indicator (RI). There are few questionnaires in each category for users to answer, and the researchers calculate the awareness level from the answers. The questions in the questionnaire are designed to test the knowledge of the users. Table 2.2 shows the sample of knowledge-based question in each focus area. Most of the questions in the questionnaire are implicit question. Meaning that this measurement method need users to answer the question, which is not suitable to adopt as the measurement method for this project.

**Table 2.2: Sample of knowledge-based question in each focus area**

Focus Area	Sample Question
Self Attitude (AT)	Personal data can be used for personal interest.
Self Behavior (BV)	Often access email from Internet cafes.
Self Cognitive (CT)	Often ask for friend's advice on computer problem
Intention to Comply (IC)	It does not matter to violate the information security rule as long as no impact at all.
Policy Compliance (PC)	It is easy to understand general written IS rule of campus.
Training Program (TP)	No possibility of occurrence on leaking answered-key in campus.
Perceived Threats (RT)	Motivation got through the awareness on the danger of negligence.
Inf. Security Awareness (IS)	Concern for the impoact will be borne for incident.
Peer Performance (PP)	Every user in campus network will obey the rule.
Social Pressure (SP)	It is common to share password with wife or close friend.
Religious Indicator (RI)	Campus provides good facilities for praying.



Velki, Solic, and Ocevcic (Velki et al., 2014a) created a measurement method called User's Information Security Awareness Questionnaire (UISAQ). The UISAQ perform the measurement by requires user answers 33 questions in the questionnaire. There is no specific focus area stated in UISAQ. Like the previous measurement, the most of questions in UISAQ are in the form of knowledge-based question.

Another systematic approach for awareness measurement was proposed by Parson, McCormac, and Butavicius (Parsons, McCormac, Butavicius, et al., 2014). The researchers developed a questionnaire called the Human Aspects of Information Security Questionnaire (HAIS-Q). HAIS-Q measures an individual's security awareness under 7 focus areas: Password management, Email use, Internet use, Social networking site use, Incident reporting, Mobile computing, and Information handling. The point schemes of the questionnaire are from 1 to 5, of which 1 represents strongly disagree and 5 represents strongly agree. The security awareness is calculated from the point of the questionnaire. The focus areas of HAIS-Q is compatible with the categories of security recommendation.

Khan and the team proposed a cyber security awareness measurement model (APAT) for awarness measurement (Khan et al., 2020). The term APAT refers to four steps of the measurement process: Analyze, Predict, Awareness, and Test. This measurement model measures the ISA with both knowledge-based and behavior-based questions. The knowledge-based questions keep updating according to the trend of cybercrime. The behaviour-based question focus on checking the screensavers, user awareness mails, reach out session, and gaming solution. The model will eventually calculates a score for the user without categorize them into levels. Although APAT model includes both knowledge-based and behavior-based questions, the result can be only calculated with both of them. The measure of knowledge-based questions and behavior-based questions work seperately. There are only 2 types of questions which are knowledge-based and behavior-based. However, the focus areas of each measurement method can be very different from each other. Table 2.3 shows the focus areas and question type of each measurement method.

**Table 2.3: Question Type and Focus Areas of each measurement**

Measurement	Focus Areas	Question Type
ISA at the Knowledge-based Institution	Self Attitude (SA), Self Behavior (BV), Self Cognitive (CT), Intention to Comply (IC), Policy Compliance (PC), Training Program (TP), Perceived Threats (RT), Inf. Security Awareness (IS), Peer Performance (PP), Social Pressure (SP), Religious Indicator (RI)	Knowledge-based
User's Information Security Awareness Questionnaire ( <b>UISAQ</b> )	33 questions with no specific focus areas	Knowledge-based
Human Aspects of Information Security Questionnaire ( <b>HAIS-Q</b> )	Password Management, Email Use, Internet Use, Social Networking Site (SNS) Use, Incident Reporting, Mobile Computing, Information Handling	Knowledge-based & Behavior-based
Cyber Security Awareness Measurement Model ( <b>APAT</b> )	Screensavers, User awareness mails, Reach out session, Gaming solution	Knowledge-based & behavior-based

## 2.5 Existing Security Awareness Measurement System

There are 4 existing systems for security awareness measurement reviewed in this research. Each work is reviewed to analyze its features and disadvantages. Table 2.4 shows the features and disadvantages of each work.

**Table 2.4: Features and Disadvantages of each existing work**

Existing Work	Features	Disadvantages
Information Security Behaviour Profiling Framework (ISBPF) for student mobile phone users  (Ngoqo & Flowerday, 2015)	<ul style="list-style-type: none"> <li>• Behavior observation and questionnaire</li> <li>• Focus areas               <ul style="list-style-type: none"> <li>○ Use of passwords</li> <li>○ Storing sensitive information</li> <li>○ Use of antivirus software</li> <li>○ Downloading files</li> <li>○ Responding to email/SMS links</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Questionnaire needed</li> <li>• No recommendations after measurement.</li> </ul>
Securing Information Sharing Through User Security Behavioral Profiling  (Fernando & Yukawa, 2014)	<ul style="list-style-type: none"> <li>• Profiling user based on behavior</li> <li>• Semi-autonomous</li> <li>• Focus areas               <ul style="list-style-type: none"> <li>○ Password security behavior</li> <li>○ Data access and backup behavior</li> <li>○ Personal observations</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Some of the observation cannot be done without human involved.</li> <li>• For example, the observations of “Personal observations” including:</li> </ul>

	<ul style="list-style-type: none"> <li>○ Information obtained through background checks</li> <li>○ Creation of user security behavioral profiles</li> <li>○ Scheduling security awareness, education and training.</li> </ul>	<ul style="list-style-type: none"> <li>○ Forgetting keycards</li> <li>○ Leaving items unattended</li> <li>○ Ambitiousness</li> <li>• Towards to management instead of self-monitoring.</li> </ul>
An information security and privacy self-assessment (ISPSA) tool for internet users (Galba et al., 2015)	<ul style="list-style-type: none"> <li>• Self-assessment.</li> <li>• Based on UISAQ.</li> <li>• Provide security advices as respond after the assessment</li> </ul>	<ul style="list-style-type: none"> <li>• Questionnaire needed</li> <li>• No Profiling</li> </ul>
Multimedia tools fo cybersecurity awareness and education (Zhang-Kennedy & Chiasson, 2021)	<ul style="list-style-type: none"> <li>• Multimedia-based (Gaming, comics, animation, etc)</li> <li>• Much interesting compare with traditional measurement methods</li> <li>• Interactable</li> </ul>	<ul style="list-style-type: none"> <li>• No suitable for automation.</li> <li>• Time-comsuming</li> <li>• No profiling</li> </ul>

Ngoqo and Flowerday (Ngoqo & Flowerday, 2015) developed an Information Security Behaviour Profiling Framework (ISBPF) for student mobile phone users. The system using questionnaire and behaviour observation with 5 focus areas including ‘Use of passwords’, ‘Storing sensitive information’, ‘Use of antivirus software’, ‘Downloading files’, and ‘Responding to email/SMS links’ to profile the users. However, the system still depend on the questionnaire to complete the profiling process.

Meanwhile, Galba et al (Galba et al., 2015) developed Information Security and Privacy Self-Assessment (ISPSA) tool based on Users' Information Security Awareness Questionnaire (UISAQ) (Velki et al., 2014b), which measures users' potentially risky behavior and users' awareness by 33 items divided into 2 scales in the questionnaire. There are 6 subareas in ISPSA including PC maintenance, security of data, usual behavior, borrowing access data, and quality of backup. Users are required to do the assessment and the result of security awareness is calculated based on the result of the assessment. The assessment questions are extracted from UISAQ, meaning that users are still required to answer the questions but in self-assessment form.

There are many tools for cyber security awareness and education using multimedia such as gaming-based, comic-based, and animation-based (Zhang-Kennedy & Chiasson, 2021). Although these tools provide a awareness result at the end, most of the tools focus on education instead of measurement. This kind of tools are more interesting compare with the other methods above especially for the young, non-expert end-users. Users can learn cyber security knowledge, trend of cybercrime, and countermeasures from the tools and eventually increase their awareness.

In this paper, HAIS-Q will be the method of security measurement. HAIS-Q has been applied to the government organization in Indonesia by Wahyudiwan and Sucahyo (Wahyudiwan et al., 2017) and in Australia by Parsons, McCormac, and Pattinson (Parsons, McCormac, Pattinson, et al., 2014). The questions in each focus area in HAIS-Q contain behaviour-based question instead of all knowledge-based question. The answer of knowledge-based question is implicit, it is hard to collect the answer by observation. The answer of behaviour-based question can be collected from the action done by user instead of using an questionnaire. On the other hand, the focus areas in HAIS-Q is match with the countermeasure of the cybercrimes. Each of the countermeasure is related to the areas stated in the HAIS-Q.

However, there are some focus areas need to be removed because it cannot apply on an individual. The detailed discussion on the focus areas which need to be removed in the next chapter.

## 2.6 Recommendation System

Recommendations should be provided to an individual after security awareness measurement as a response, and to increase the individual's security awareness. However, people reject to act accordingly to the recommendation when they felt it is complicated. The problem is, the complexity of recommendation can be subjective. That's why the recommendation system is introduced.

Recommendation system has been used in many areas such as online shopping, movie, video, music, and others. The recommendation system is able to provide recommendations to the user based on their preferences or other users' preferences. Hence, it is able to predict which recommendation should be provided to a certain individual with suitable complexity. There are four types of recommendation systems which are collaborative, content-based, knowledge-based, and hybrid.

### 2.6.1 Collaborative Filtering

Collaborative Filtering system holds a database that contains information of user's preferences of items. In order to provide recommendations to the target user, collaborative filtering compares the preferences of the target user with other users in the database and then finds those strongly correlate with the target user. Items recommended to the target user are rated highly by those similar user collaborative filtering found in the database. To complete the whole process, collaborative filtering needs both user data and item data.

Pearson's correlation coefficient is one of the most common methods to calculate a rating (Lyons, 2014). It assigns a value from -1 to 1 to the users, and 1 represents a very strong positive correlation and -1 represents a very strong negative correlation. The users who get high positive correlation values are those users who have similar preferences in items. The Equation (2.1) to calculate the similarity between users shown below:

$$similarity(a, b) = \frac{\sum_{i \in I} (r_{a,i} - \bar{r}_a)(r_{b,i} - \bar{r}_b)}{\sqrt{\sum_{i \in I} (r_{a,i} - \bar{r}_a)^2} \sqrt{\sum_{i \in I} (r_{b,i} - \bar{r}_b)^2}} \quad (2.1)$$

From the equation, the value of similarity between "user a" and "user b" is the sum of each item 'i' in the set of item I. The symbol  $r_{a,i}$  refers to the rating of item 'i' by the user a, same goes to  $r_{b,i}$ , and the symbol  $\bar{r}_a$  refers to the average rating of the user a. The

users with high similarity value considered as ‘neighbor’ in the matrix. With the similarity value calculated from the Pearson’s correlation coefficient and N nearest neighbor, a prediction value to determine whether a user like an item can be calculated as the Equation (2.2) below:

$$prediction(a, i) = \bar{r}_a + \frac{\sum_{b \in N} similarity(a, b) * (r_{b, i} - \bar{r}_b)}{\sum_{b \in N} similarity(a, b)} \quad (2.2)$$

When there is a lack of ratings between users and items, it will lead to a problem known as sparsity. It happened because the ratings between users and items are not enough for the recommendation system to make an accurate prediction. Another problem may occur in collaborative filtering known as cold start. It happens when a new user has only a few ratings and the recommendation system is not able to provide an accurate prediction based on the few ratings only.

## 2.6.2 Content-based Filtering

Content Based filtering needs only item data. Content-based filtering holds a profile for each item. Content-based filtering provides recommendations by comparing the attributes or features between items, and then choose the most similar item to recommend. However, some content-based filtering systems build both item profile and user profile. Item profile contains attributes of items while user profile contains user preferences, view history, rating, and more (Li et al., 2019). Content-based filtering needs attributes of one item in order to compare the similarity to another item before it can do the prediction. However, collecting the attributes about items will be difficult if those attributes are qualitative.

Content-based filtering first checks the words in meta-data of the items and then creates a vector with value 1 to represent if a word exists in the meta-data, and 0 represent if a word does not exist (Lyons, 2014). Then, it compares the vector with another vector from other items to determine the similarity. This approach has a defect, the vector does not consider the frequency and weight of words in the meta-data. Term Frequency-Inverse Document Frequency (TF-IDF) is a technique to resolve this problem. TF-IDF considers the frequency of the words that appear in meta-data, and also the weight of the words (Salton et al., 1975).

Content-based filtering calculates the similarity of items by analyzing the words inside meta-data, and then select those items nearest to the item in which the user is interested. However, some items are hard to describe or express in words, the features will be hard to extract for non-text-based items, especially if it is more subjective. Another issue of content-based filtering is the “Filter Bubble” (Yao et al., 2018), which means the content-based filtering will recommend similar items, and only similar items to the user. This makes the recommended items are always from the same category, and the users cannot reach other items outside the “bubble”. Lastly, content-based filtering also has the cold start problem. It is hard to provide recommendations to a new user without the user bought anything before.

### **2.6.3 Knowledge-based**

Knowledge-based recommendation systems use the detailed rule of the problem domain and items’ attributes for predictions and provide recommendations to users. Different from collaborative and content-based filtering, a knowledge-based recommendation system does not collect information about user ratings and history, it collects specific requirements from users to provide the items which users may like (Lyons, 2014).

The case-based reasoning (CBR) models the products and then searching the similar product with the requirements partially described by the user (Lorenzi & Ricci, 2003). For example, a customer is looking for a product. The user makes some explicit requirements about the product, then CBR starts to search the case base for the products which meet those requirements. A set of products will be recommended to the user based on the requirements, the user can modify the requirements if he/she is not satisfied with those recommendations to get a set of new recommendations.

Utility-based recommender systems gathering the interest level or weight of a user has in a specific attribute and then search the items that meet the overall utility of the user. The total utility value is the sum of all the values of the item, multiplied with the similarity function used in case-based recommendations (Lyons, 2014). The utility values represent a ranking of items based on the similarity level between the items and the requirements set by the user.



#### **2.6.4 Hybrid Recommendation System**

A hybrid recommender system is a combined recommender system that may include any two approaches between collaborative, content-based, and knowledge-based, or all of them. A hybrid recommender system takes advantage of the strengths of each combined approach. By combining each approach, the result of one algorithm can be feed into the input of the second algorithm or used in a parallel way (Zanker, 2010).

Collaborative and content-based recommendation systems have the same problem which is the sparsity problem, both collaborative and content-based recommenders are performing well with a high density of information. While the knowledge-based recommendation system does not influence by sparsity problem since it focuses on the problem domain rather than the users and items domain. However, the knowledge-based recommendation system is not good at mapping users and items. Looking back at collaborative filtering and content-based filtering, the reason which causes the sparsity problem, dependencies of users and items, could help to develop an association between items and users. Besides, the collaborative recommendation system can handle well the weakness of content-based recommenders such as dealing with the items which are difficult to exact attributes. Similarly, content-based recommenders can provide accurate recommendations with very few user-item ratings (Lyons, 2014). Therefore, the weaknesses of each approach can be solved by combining the approaches, using the strength of each approach to cover it.

#### **2.7 Recommendation**

There are lots of sources that have provided recommendations in cybersecurity domain including but not limited to journals, web sources, companies, and government organizations. The range the recommendations for the cybersecurity domain can be very wide, including many of the areas such as password, network, data access control, activity monitoring, backup policy, and so on. The range of depth of each area can also be very wide. For example, in the network area, the recommendation can be very general like suggests user to install a firewall, also can be very professional which requires professional knowledge like the configuration of an Intrusion Detection System.

Many web sources provide the best practices for network security management including establishing information security framework, data access control, monitoring user activity, giving training to employees, and so on (Donovan, 2017; Juno Risk Solutions, 2015; McCarthy et al., 2014; WaterISAC, 2015).

The United States Computer Emergency Readiness Team (US-CERT) provide general cybersecurity recommendation in areas including threats, email and communication, general information, general security information, mobile devices, privacy, safe browsing, software and applications, network defense and enterprise security, and archive. The recommendations have been well categorized, people can find their needs easily with the categories. Each category has few sub-topics. There are clear explanations for each sub-topics including definition, works theory, deliver method, protection, prevention, and response. However, larger coverage of recommendations in each category could lead people to giveup as it is complicated.

Malaysia government organization Cyber Security Awareness for Everyone (CyberSAFE) also provide general cybersecurity recommendation on their website. The areas including emails and spam, virus and worms, password protection, and so on. The recommendations are categorized but there is no sub-topics and explanations in each category. No extra information provide besides the recommendation.

Another Malaysia government organization Malaysia Computer Emergency Response Team (MyCERT) also provide alert and advisories on their website. The organization keep updating the latest security information and list out the recents threats, security updates, and reports on the webpage. However, the alerts and advisories are not categorized. It is hard for people to find the information they need.

The focus areas of recommendations from different sources are not the same. But there are still some of the common focus areas which often appear. Also, some of the recommendations are more suitable to apply to an organization instead of an individual. In this paper, the recommendations will be selected from CyberSAFE and US-CERT as the recommendations are general and suitable to apply to an individual.

## **2.8 Summary**

There are 8 cybercrimes listed by MyCERT including content-related, cyber harassment, denial-of-service, fraud, intrusion, intrusion attempt, malicious code, and spam. Among them, fraud, intrusion, and malicious code account for the majority. People should pay more attention to these 3 types of cybercrimes. Security awareness is reflected in the behavior of the people in the cybersecurity domain. Measuring security awareness using information collected from people's security behavior is feasible. There are 3 ISA measurement methods reviewed including ISA at the knowledge-based institution, UISAQ, and HAIS-Q, and the HAIS-Q is adopted for this paper. Using questionnaire and not-atumated are the common disadvantages of the existing systems. There are 4 types of recommendation systems including collaborative filtering, content-based filtering, knowledge-based filtering, and hybrid. Collaborative filtering classifies users into groups, while content-based filtering finds the similarity using items attributes. Knowledge-based filtering provides recommendations based on user's requirements. A hybrid recommendation system combines multiple filtering techniques in order to cover the weakness of each other. Furthermore, the range of the cyber security recommendations can be very wide, it can be very general or very professional. The CyberSAFE provides very general categorized recommendations that lack of functionality while MyCERT provides better recommendations but uncategorized. Meanwhile, US-CERT provides detailed and categorized recommendations.

## **CHAPTER 3**

### **METHODOLOGY**

#### **3.1 Overview**

In this chapter, the methodologies that aim to solve the problems of low effectivity of security advice due to ambiguity and complexity, and lack of ISA measurement for individual. This paper proposed an ISA profiling and recommendation system that using customized HAIS-Q to suite individual ISA measurement and generate ISA profile to train the collaborative recommendation system. The recommendation system adopt matrix factorization and RNN-LSTM algorithm to train the model. The processes of this research are divided into 6 parts:

- Parameter compositions
- Security awareness measurement
- Recommendation selection
- Data collection
- Modeling
- Evaluation

Parameter compositions apply 5 focus areas from HAIS-Q: Password management, email use, internet use, social network site (SNS) use, and information handling. Each area contains a checklist of security status that will be collected from an individual's computer which is related to 8 cybercrimes in Malaysia.

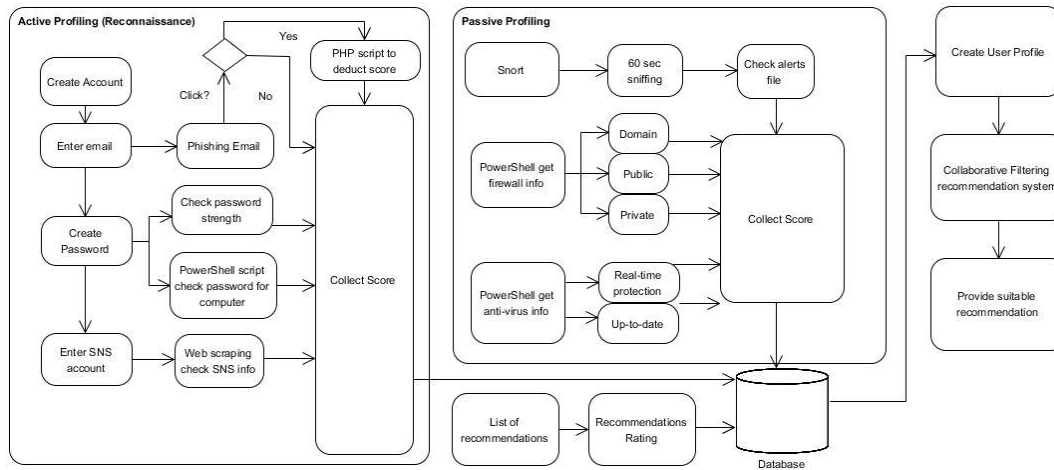
Security awareness measurement refers to the calculation of security awareness. Security awareness is calculated by using the combination of focus areas and the weight of cybercrimes. The level of security awareness is divided into 3 levels: Good, average, and poor. Each user will be assigned the level after security awareness is measured.

Recommendation selection refers to the process of selecting recommendations that are suitable for this research. The recommendations that are related to the focus areas and can be applied to an individual will be selected.

Data collection refers to the benchmark data collection. Benchmark data collection is done by distributing a script to the public. The script will first measure the user's security awareness level, and then provide recommendations by using knowledge-based filtering. Users are required to rate the recommendation whether it is suitable for them. The rating will become benchmarks to train collaborative filtering.

Modeling refers to the method of building a model for the collaborative filtering recommendation system. The library to build the collaborative filtering recommendation system is using Fastai.

Evaluation refers to the evaluation method used to determine the reliability of the recommendation system. RMSE and confusion matrix are the main evaluation methods to examine the recommendation system. The overview of system design is shown in Figure 3.1.



**Figure 0.1: System design overview**

The users are required to enter their email address, SNS account, and create a new password for the first time. A phishing email will be sent to the entered email address for the phishing test. The SNS account of the user will be analyzed by a web scraping script. The pattern of the password will be checked to analyze the strength. In the meanwhile, the passive profiling analyzes the network of the user by using Snort, analyzes firewall and anti-virus status by the PowerShell scripts. After the steps above, the system will generate user profile for the user and turn the collected information into a score and then upload into the database. The profiles will be the input for the recommendation system to train the model. The details of the processes are discussed in following sections.

### 3.2 Parameter composition

H AIS-Q is originally designed to measure employees' security awareness. This research is aiming to measure an individual's security awareness level and provide recommendations by collecting information from personal computers without answering a question. There are some unnecessary focus areas that need to be removed from H AIS-Q. There are 2 focus areas that need to be removed from H AIS-Q. The first is 'Incident

reporting’, this focus area describes an incident report in an organization such as reporting bad behavior by colleagues, which is an unnecessary area for an individual. The second focus area needed to be removed is ‘Mobile computing, this focus area describes the network access control inside an organization such as access work email using a public network, which is also an unnecessary area for an individual.

The parameters to measure security awareness are derived from the 5 focus areas in HAIS-Q. This includes *Password management, email use, internet use, SNS use, and information handling*. Table 3.1 shows the focus area and sub-areas.

**Table 3.1: Focus Areas of modified HAIS-Q**

Focus Area	Sub-Areas
Password Management	<ul style="list-style-type: none"> <li>• Locking computer</li> <li>• Password strength</li> </ul>
Email Use	<ul style="list-style-type: none"> <li>• Opening attachments or links</li> </ul>
Internet Use	<ul style="list-style-type: none"> <li>• Monitoring network traffic</li> </ul>
Social networking site (SNS) use	<ul style="list-style-type: none"> <li>• Posting sensitive information on SNS</li> </ul>
Information Handling	<ul style="list-style-type: none"> <li>• Firewall</li> <li>• Anti-Virus</li> </ul>

Each area contains a checklist of security status that are collected from individual’s computer that are related to the 8 most common cybercrimes in Malaysia. In this research, the intrusion and intrusion attempt are combined into 1 category. The focus area which is considered as the area of countermeasure for a certain cybercrime will be selected as the measures focus area for the cybercrime. For example, the countermeasure of fraud including be aware of a phishing email, and do not share personal detail on SNS. Then, the focus area of fraud will be email use and SNS use. Table 3.2 shows the countermeasures for each cybercrime.

**Table 3.2: References and countermeasure of cybercrime**

Cybercrime	Countermeasures	Reference
Content related	<ul style="list-style-type: none"> <li>• Install anti-virus</li> <li>• Do not send a picture of yourself when chatting</li> </ul>	CyberSAFE
Cyber Harassment	<ul style="list-style-type: none"> <li>• Do not flirt online</li> <li>• Choose a genderless screen name when chatting</li> </ul>	CyberSAFE
Denial-of-Service	<ul style="list-style-type: none"> <li>• Turn on firewall</li> </ul>	US-CERT
Fraud	<ul style="list-style-type: none"> <li>• Be aware of phishing email</li> <li>• Do not share personal detail on SNS</li> </ul>	CyberSAFE
Intrusion / Intrusion Attempt	<ul style="list-style-type: none"> <li>• Authentication</li> <li>• Aware of email attachment</li> <li>• Turn on firewall</li> <li>• Install anti-virus</li> </ul>	US-CERT, CyberSAFE, (Litoussi et al., 2020)
Malicious Code	<ul style="list-style-type: none"> <li>• Never open e-mail attachment from stranger</li> <li>• Enable firewall</li> <li>• Install anti-virus software</li> <li>• Do not download pirated or cracked program</li> </ul>	CyberSAFE, US-CERT
Spam	<ul style="list-style-type: none"> <li>• Never post email address publicly</li> <li>• Do not reply spam</li> <li>• Download spam filtering tools and anti-virus</li> </ul>	CSA Singapore (Cyber Security Awareness Alliance, n.d.)

For localized contexts, each cybercrime are weighted based on the average proportion of cyber-incidents that are reported in Malaysia from the year 2017 to 2019. Table 3.3 shows the mapping of cybercrimes to their corresponding focus areas.

**Table 3.3: Relationship between cybercrimes, focus area, and weight**

<b>Cybercrime</b>	<b>Focus Area</b>	<b>Weight</b>
Content related	<ul style="list-style-type: none"> <li>• Information Handling</li> <li>• SNS Use</li> </ul>	0.023
Cyber Harassment	<ul style="list-style-type: none"> <li>• Information Handling</li> <li>• SNS Use</li> </ul>	0.024
Denial-of-Service	<ul style="list-style-type: none"> <li>• Information Handling</li> </ul>	0.002
Fraud	<ul style="list-style-type: none"> <li>• Email Use</li> <li>• SNS Use</li> </ul>	0.732
Intrusion / Intrusion Attempt	<ul style="list-style-type: none"> <li>• Password Management</li> <li>• Internet Use</li> <li>• Email Use</li> <li>• Information Handling</li> </ul>	0.138
Malicious Code	<ul style="list-style-type: none"> <li>• Internet Use</li> <li>• Email Use</li> <li>• Information Handling</li> </ul>	0.069
Spam	<ul style="list-style-type: none"> <li>• Email Use</li> <li>• Information Handling</li> </ul>	0.012

### **3.3 Security awareness measurement**

Security awareness measurement is performed by checking every sub-areas in each focus area in the user's computer. The score which represents security awareness level will



be calculated according to the result. The users will be categorized into 3 levels based on the ISA result, the levels including *poor* (1), *average* (2), and *good* (3) as shown in Table 3.5. Equation (3.1) shows the calculation formula of security awareness measurement is as follow. The equation define the weight of each cybercrime by calculate the percentage of the specific cybercrime across the total cases. The score of each focus area is defined as the sum of the weight of the cybercrime divided by the fraction of the focus area related to the cybercrime. The definition of each variable in the equation is shown in Table 3.4.

$C = \{\text{Content related, Cyber harassment, DDoS, Fraud, Intrusion, Malicious Code, Spam}\}$

$I = \{\text{Email use, Password management, SNS use, Internet use, Information handling}\}$

$$c \in C, w_c = \frac{\text{cases}_c}{\text{Total cases}}$$

$$\sum_{i \in I} \text{score}_i = \frac{w_c}{r_i} * n_i \quad (3.1)$$

**Table 3.4: Equation variables definition**

Variable	Definition
C	The set of cybercrime
c	Cybercrime in the set of cybercrime
w	Weight
I	The set of focus area
i	Focus area in the set of focus area
r	Total number of focus areas
n	Number of focus area categorized and related to the cybercrime

**Table 3.5: Range of security awareness level**

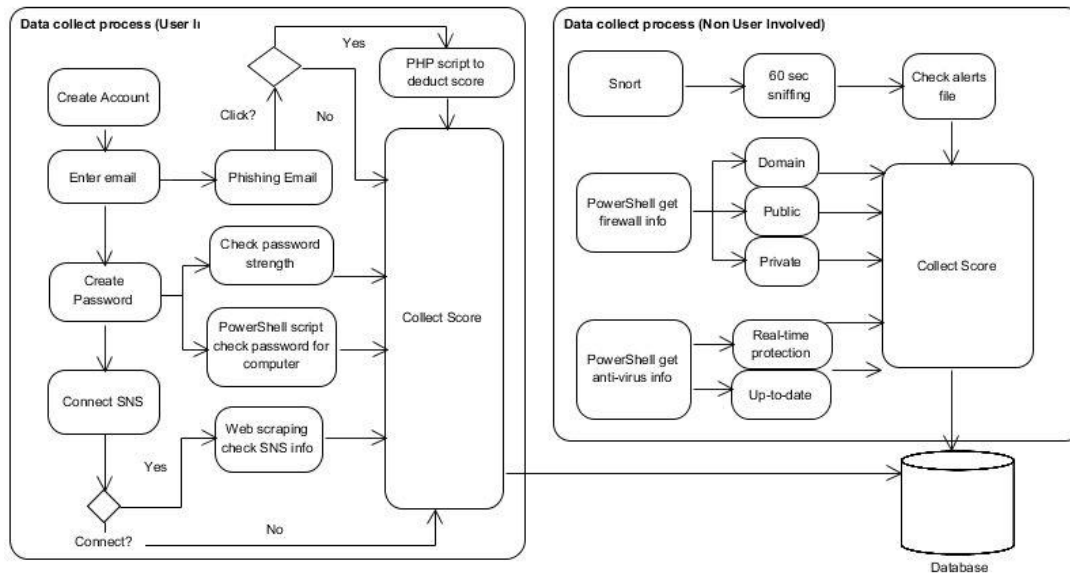
<b>Level</b>	<b>Score (%)</b>
Good (3)	80.00 – 100.00
Average (2)	60.00 – 79.99
Poor (1)	0 – 59.99

### **3.4 Recommendation selection**

The recommendations are selected from US-CERT that related to the focus area accordingly. There are 2 types of recommendation: preliminary and trigger. The preliminary recommendation is provided after security awareness measurement is done. It is to provide extra information to the user in order to improve security awareness. Trigger recommendation is provided when there is a vulnerability found after checking the sub-areas in each focus area. It is to point out the issue and also provide recommendations for the issue. The focus area of Information Handling contains 6 recommendations, 3 for firewall and 3 for anti-virus. Each of the rest of the focus areas contains 3 recommendations.

### 3.5 Data collection

A set of 101 users is selected for the CA profiling; this fulfills the requirement of Slovin's Formula with a confident level of 90% and margin-of-error at 10% for the sample selection. The profiling process is entirely autonomous; it is handled by a Python script that runs in the background while monitoring user actions. Initial security recommendations prompted the users when some specific unsafe actions performed triggers these alarms. The users are asked to rate these recommendations on a scale of 1-5, depending on the user-friendliness of these suggestions. This profiling module is hosted on-premise that runs on myPHP 4.9.5; meanwhile, all users' profiles are stored in MariaDB 10.3.16. The script continuously checks for actions that raise red flags. Figure 3.2 shows the data collection program design.



**Figure 0.2: Flow diagram of data collection**

The data collection processes of each focus area are as follow:

- *Password management*

This focus area is divided into 2 parts: locking computer, and password strength. The checking process of sub-area 'Locking Computer' is performed by a PowerShell script which checks whether the user set a password lock for the computer. Due to privacy issues, looking for the password inside the computer is prohibited. Alternately, the user is required to create a new password, the script analyzes the created password for password strength. The definition of password

strength is defined differently by different organizations. This research adopt Microsoft password practices recommendations. Table 3.6 shows the relationship between sub-area and checklist of password management.

**Table 3.6: Sub-area and checklist of password management**

Sub-area	Checklist
Locking computer	Set password for the computer
Password strength	At least 8 characters
	Contains upper case
	Contains lower case
	Contains symbol
	Contains numbers

- *Email use*

A cron-job, scheduled job that run on fixed times is implemented to periodically send curated phishing emails to target users without prior notifications. The email contents vary from extremely overt to extremely indistinguishable. The intuition is that 'this test reveals the parity of phishing awareness depending on what the user clicks at and falls for.' The score will be deducted if the user clicks on the link attached to the phishing email. As the user clicks the link in the phishing email, the user's email address is sent to the PHP web page. The webpage will then locate the user using the email in the MariaDB and recalibrate the scores.

- *Internet use*

Snort (network sniffer) is integrated inside the script to periodically check for possible violations and triggers at every 60 seconds interval. The snort is configured with the general network rules defined by the Snort community. These rules consist of suspicious network traffic, including known malware traffic like ICMP flooding, DNS poisoning, FTP redirection, and so on.

- *SNS use*

This process is performed by a web scraping script using Python programming language. It crawls the HTML element from the user's Facebook page and searches for sensitive information like phone numbers and addresses. Users are systematically penalized whenever they are found to be over-sharing sensitive information.

- *Information handling*

The script is programmed to check for current firewalls and antivirus status, patches, and misconfigurations. Using 'netsh advfirewall show all profiles state', the script automatically detects the Windows Firewall settings that come standard in all Windows-based machines. Users ISA is adjusted accordingly depending on the types of Firewall that are disabled, like 'domain', 'public' or 'private'. Meanwhile, a custom crawler reads antivirus implementations using registry-hacks to detect the current antivirus protection that users set had on the machine. Table 3.7 shows the sub-area and checklist of information handling.

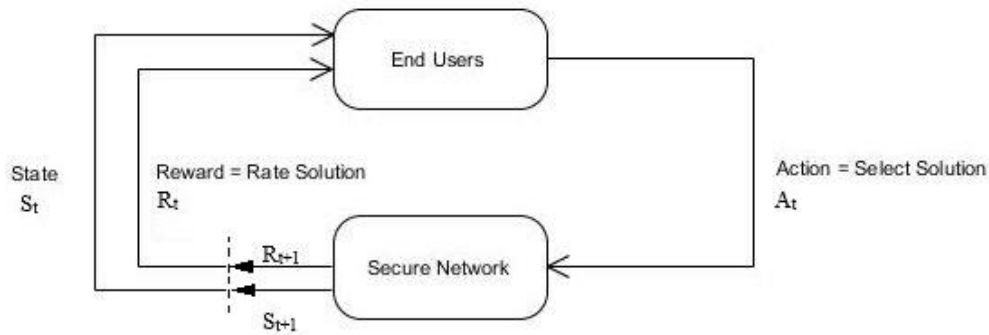
**Table 3.7: Checklist of information handling**

Sub-area	Checklist
Firewall	Domain
	Private
	Public
Antivirus	Real-time protection
	Up-to-date

### 3.6 Modeling

There are two recommendation models trained using collaborative filtering on Fastai/Pytorch: ISA-aware recommendation and non-ISA aware recommendations. In this

section, the non-ISA aware recommendations is first discussed. The research hypothesize that a user-centric recommendations can be modeled using user ratings for existing security solutions as the training data. Figure 3.3 shows the intuition of reinforcement learning to upvote useful suggestions and downvote some least useful ones based on user feedbacks.



**Figure 0.3: Reinforcement learning using user ratings on security recommendations**

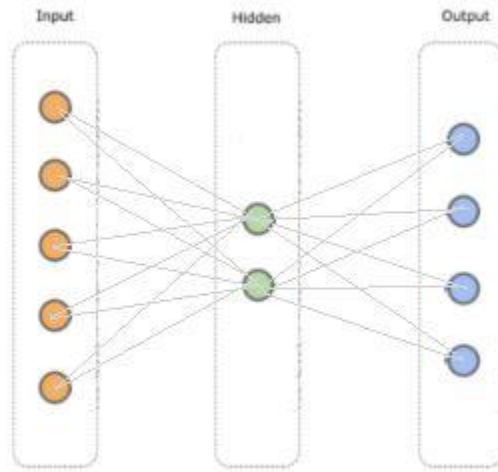
The crowd-sourced ratings for the set of security advisories to train models that are users-driven using reinforcement learning. The users are first asked to rate the first level recommendation from the preliminary rounds that is later used to train the triggered recommendations. Each of the suggested solutions are ranked from *level 1-5* based on ‘ease of uses’, ‘feasibility’ and ‘user-friendliness’. These labels (Table 3.8) are then used to train the triggered recommendation using reinforcement learning; such that a highly rated solutions is rewarded highly while a poorly rated solutions is penalized towards finding a global optimum solutions.

**Table 3.8: Snippets of users-labelled dataset (rating)**

	<b>R1</b>	<b>R2</b>	<b>R3</b>	<b>R4</b>	<b>R5</b>	<b>...</b>	<b>R20</b>
<b>User 1</b>	1	Null	2	3	4	...	5
<b>User 2</b>	3	2	3	3	2	...	5
<b>User 3</b>	3	3	Null	3	5	...	3
<b>User 4</b>	Null	2	3	Null	3	...	4

<b>User 5</b>	4	3	Null	5	2	...	5
---------------	---	---	------	---	---	-----	---

The recommender is trained using a factorization machine without biases and identity activation functions. Matrix factorization is a method to generate latent features by multiplying two entities. The factorization machine's NN architecture is visualized in Figure 3.4. The neural network (NN) is designed in 3-layers; start with an input layer  $L^1$  with  $N$  inputs, a hidden layer  $L^2$  with  $K$  units and an output layer  $L^3$  with  $M$  units. The size of the hidden layer determines the dimension of the latent factors. Matrix  $P$  represents the relationship between a user and the features. Matrix  $Q$  represents the relationship between an item and the features. The prediction of a rating of an item can be generated by the calculation of the dot product of user and item.



**Figure 0.4: 3-layers neural network**

The Matrix factorization generate the latent features by define a set of users ( $U$ ), items ( $D$ ),  $R$  size of  $|U|$ , and  $|D|$ . All the rating given by users are includes in matrix  $|U| * |D|$ . The purpose is to find out the latent features ( $K$ ). Given with the matrices  $P=(|U|*K)$  and  $Q=(|D|*K)$ , the result  $R$  can be calculated using Equation (3.2).

$$R \approx P * Q^T = \hat{R} \quad (3.2)$$

It is hypothesized that the preferred security recommendations are highly similar for the group of people with similar cyber-security awareness. There are two models: (a) 'non-ISA aware' model (model\_1) and (b) 'ISA-aware model' (model\_2). Model\_1 is

trained without ISA context using dataset in Table 3.9 that contains unsorted user ratings for various security recommendations. Meanwhile, *Model\_2* is trained using dataset in Table 3.10 that contains ratings from end users who are classified based their ISA levels. The hyperparameter tuning for these models are automated using PyTorch AX, the parameters of learning rate=5e-3, batch size=2 and weight decay=0.1 are selected with early stopping.

**Table 3.9: Unsorted user ratings for various security recommendations**

User	password	network	R1	R2	...	R20
User 1	0	0	1	Null	...	5
User 2	0	0	3	2	...	5
User 3	1	0	3	3	...	3

**Table 3.10: Ratings from end-users who are classified based on their ISA levels**

User	password	network	R1	R2	...	R20	Score	Level
User 1	0	0	1	Null	...	5	0.7192	2
User 2	0	0	3	2	...	5	0.7958	2
User 3	1	0	3	3	...	3	0.8235	3

The recommendation system adopt Recurrent Neural Network (RNN) for the modeling. RNN is a type of neural network that modeling the sequential data. RNN able to predict the next outcome based on the previous input. Since the user profiles and the ratings are keep updating time-to-time, it become the sequential data. RNN helps to generate prediction that is more accurate by modeling the user profiles and ratings as sequential data. Long Short-Term Memory network (LSTM) is an extension for RNN. LSTM extend the memory of RNN so that it can remember longer period of inputs. Other than that, LSTM



also has gates to control on new input entering, unimportant node removing, and the impact of the input to the output. With the RNN-LSTM, the model is able to handle the data in a wide range of time while improving the result.

### 3.7 Evaluation

The model is evaluated using Root Mean Square Error (RMSE) and a confusion matrix. RMSE is a popular method that is used to measure the error of a recommendation system. The value of RMSE can tell how far the predicted rating is from the actual rating. RMSE is calculated using Equation (3.3).

$$RMSE = \sqrt{\frac{\sum(x-x_1)^2}{n}} \quad (3.3)$$

Other than RMSE, confusion matrix determine the accuracy of a recommendation system by classifying the result into 4 categories: True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN) as shown in Table 3.11. After confusion matrix is generated, the accuracy of the recommendation system can be determined by using Equation (3.4).

- **True-Positive:** The model predicts the user prefers the recommendation and the user actually prefers the recommendation.
- **True-Negative:** The model predicts the user prefers the recommendation but the user actually not prefers the recommendation.
- **False-Positive:** The model predicts user does not prefer the recommendation but the user actually prefers the recommendation.
- **False-Negative:** The model predicts user does not prefer the recommendation and the user actually not prefers the recommendation.

**Table 3.11: Confusion matrix**

	<b>Recommended</b>	<b>Not recommended</b>
<b>Preferred</b>	True-Positive	False-Negative
<b>Not preferred</b>	False-Positive	True-Negative

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.4)$$

### 3.8 Summary

In this chapter, the design of the data collection process, the method to select recommendation, the design of the recommendation system, and the evaluation method were described in detail. There are 2 focus areas removed from HAIS-Q because the areas are focus on employees and organization. Each cybercrime is addressed by one or more focus areas. The method of security awareness measurement was described in detail including the calculation and leveling range. The processes of data collection were broken into parts. The method of collection process of each focus area was described in detail. The recommendations in this research are selected from US-CERT that related to the focus areas. A model-based collaborative filtering recommendation system is used to examine the hypothesis proposed. RMSE and accuracy using confusion matrix are the main evaluation method to examine the accuracy of the model.

## CHAPTER 4

### RESULTS

#### 4.1 Overview

In this chapter, the results of data collection and collaborative filtering are presented. Start with the statistical analysis of collected data, followed by the training result of the model, and closing with the implications of the experiment's results.

#### 4.2 Data collection

Table 4.1 showed the experimental results for the first round of ISA profiling and second round of ISA profiling. In the first profiling, most of the users met the stringent **password management** requirements at 87.13%. This indicates that users are well educated in setting strong passwords as the first-level authentication. Note that most modern apps show the password guideline to meet minimum security requirements to restrict weak passwords. Meanwhile, only 7.92% of users fall victim to **phishing emails** when presented with emails containing malicious links. This showed that most users are cautious to external links despite these phishing emails are well crafted for click-

baits, like using URLs that closely resemble the original links or special incentives to attract users. It is worth noting that some email clients like Google Mail might have already filtered some of these emails (like Google Anti-spam) before they are presented to the users. In the **SNS use**, it is found that only 6.93% of users overshare their social networks. This is comparably lower than other focus areas, mainly because modern apps are becoming more privacy-focused. For example, iPhone users are now warned with a prompt of apps tracking and the types of information being shared by the apps they used in iOS14.6. This implies that the improved user awareness of cyber-threats in recent days is driven by built-in intelligence at the devices and operating systems levels. Lastly, most users passed the set of tests for **Information Handling** since modern OS like Windows 10 and MAC OS automatically push security updates and patches to the end-users and optimize security configurations in real-time. On second profiling, the ISA result is slightly better than the result from first round of profiling. This is because the users realized their weaknesses through the profiling system, and then take action accordingly.

The result conclude that the average ISA among the respondents is above average. The profiling system can help to increase the ISA of users. However, it do not distinguish if the improved ISA results from users are getting more educated or the security features built into modern apps and devices.

**Table 4.1: Result of data collection**

Focus area	Sub-area		Frequency	Percentage	Frequency	Percentage
			(1st)	(1st)	(2nd)	(2nd)
Password Management	Locking computer		20	19.80%	22	21.78%
	Password Strength	At least 8 characters	88	87.13%	92	91.09%
		upper case	24	23.76%	27	26.73%
		lower case	32	31.68%	39	38.61%
		Contains symbol	54	53.47%	56	55.45%

		Contains numbers	53	52.48%	60	59.41%
Email use	Phishing test		8	7.92%	1	0.99%
Internet use	Network traffic		5	4.95%	5	4.95%
SNS use	Sensitive Information on SNS		7	6.93%	6	5.94%
Information Handling	Anti-virus	Real-Time protection	97	96.04%	101	100%
		Up-to-Date	97	96.04%	101	100%
	Firewall	Domain	97	96.04%	101	100%
		Private	97	96.04%	101	100%
		Public	97	96.04%	101	100%

The automated ISA measurement is implemented successfully by satisfying each focus area and sub-area with the data captured, representing respondents' behavior from their computer. The ISA is measured without any single questionnaire and applicable to the individual. The user profile is composed of data captured from each focus area and its sub-area. The profile exposes the security vulnerabilities and provides valuable insight into which area needs to be improved. Statistical analysis can be implemented using data collection to identify the weak focus area of the majority. Other than that, the profile calculated and categorized respondents into three levels. The results of ISA level reflect what people should take actions at different stages. Our results showed that there 15 respondents are categorized as poor ISA level, 78 average, and 8 good. Only 7.92% of respondents considered having high ISA. The rest of 92.08% of respondents need to improve their ISA. The result of the ISA level of respondents is shown in Table 4.2. Hence,

the Malaysian government can exploit ISA profiling of the public and provide a suitable method to increase public ISA.

**Table 4.2: Table of user security awareness level**

Category	Frequency (1st)	Frequency (2nd)	Score range	Category
Good (3)	8	12	0.8 – 1.0	Good
Average (2)	78	80	0.6 – 0.7999	Average
Poor (1)	15	9	0 – 0.5999	Poor

The recommendations are rated by the respondents. The rating is using point schemes as 1 which represents ‘very unsuitable for me’ to 5 which represents ‘very suitable for me’. The recommendations that aim to resolve the existing issue instead of provide general information are accepted by users as they have a good average rating. This also meaning that the problem of ambiguous recommendation can be eliminated with trigger recommendation. The ambiguity of the recommendation comes from the way of providing the recommendation. Listing all the recommendations at once confuse the people as there are too many advices and people do not know which advice they should apply. The trigger recommendation points out the existing security vulnerabilities that found in user’s computer while provide security advice to the user. The average rating of trigger recommendations in each focus area is shown in Table 4.3. The average rating of trigger recommendations in each focus area is above 4, which means users are satisfied with it.

**Table 4.3: Average rating of each trigger recommendation**

	Password Management	Email Use	SNS Use	Internet Use	Information Handling
Average Rating	4.4	4.7	4.5	4.1	4.2

### 4.3 Training Results

The reliability of the recommendation is determined by using the RMSE metric and confusion metric. RMSE is calculated by comparing the predicted value with the true value. The confusion metric determines the precision and recall of the model. The model is trained with parameters shown in Table 4.4 while batch size defined the number of training samples utilized in one iteration, epoch defined the number of passes of the dataset to the model, learning rate defined the step size at each iteration while moving toward a minimum of a loss function, and weight decay defined whether the model is going to be underfitting, just right, or overfitting.

**Table 4.4: Parameters of model training**

Parameter	Value
Batch size	2, 4, 8, 16, 32, 64
Epoch	10
Learning rate	5e-4, 1e-3, 5e-3, 1e-2, 5e-2
Weight decay	0.001, 0.01, 0.1, 1, 10

Each combination of parameters are tested in order to find out the best parameters setting for model training. After training with each combination of the parameters, the best setting for this model is using batch size with 2, learning rate with 5e-3, and weight decay with 0.01. The RMSE result of each combination is shown in Table 4.5 to Table 4.9. The smaller the RMSE, the better the result.

**Table 4.5: RMSE result of each batch size and learning rate with weight decay value 0.001**

RMSE Table	BS=64	BS=32	BS=16	BS=8	BS=4	BS=2
LR = 5e-4	1.40	1.40	1.33	1.26	1.22	1.16

<b>LR = 1e-3</b>	1.35	1.36	1.33	1.17	1.11	0.96
<b>LR = 5e-3</b>	1.10	0.85	0.74	0.74	0.74	0.67
<b>LR = 1e-2</b>	0.88	0.70	0.70	0.69	0.70	0.78
<b>LR = 5e-2</b>	0.78	0.82	0.84	0.78	0.70	0.79

**Table 4.6: RMSE result of each batch size and learning rate with weight decay value 0.01**

<b>RMSE Table</b>	<b>BS=64</b>	<b>BS=32</b>	<b>BS=16</b>	<b>BS=8</b>	<b>BS=4</b>	<b>BS=2</b>
<b>LR = 5e-4</b>	1.58	1.53	1.44	1.32	1.22	1.00
<b>LR = 1e-3</b>	1.43	1.32	1.12	1.00	0.79	0.80
<b>LR = 5e-3</b>	0.85	0.80	0.76	0.74	0.69	0.61
<b>LR = 1e-2</b>	0.74	0.67	0.74	0.73	0.68	0.78
<b>LR = 5e-2</b>	0.78	0.76	0.84	0.84	0.72	0.73

**Table 4.7: RMSE result of each batch size and learning rate with weight decay value 0.1**

<b>RMSE Table</b>	<b>BS=64</b>	<b>BS=32</b>	<b>BS=16</b>	<b>BS=8</b>	<b>BS=4</b>	<b>BS=2</b>
<b>LR = 5e-4</b>	1.46	1.42	1.37	1.36	1.32	1.22
<b>LR = 1e-3</b>	1.40	1.28	1.27	1.23	1.09	0.96
<b>LR = 5e-3</b>	1.13	0.94	0.76	0.76	0.69	0.74
<b>LR = 1e-2</b>	0.78	0.75	0.80	0.69	0.68	0.68

<b>LR = 5e-2</b>	0.78	0.72	0.72	0.75	0.69	0.67
------------------	------	------	------	------	------	------

**Table 4.8: RMSE result of each batch size and learning rate with weight decay value 1**

<b>RMSE Table</b>	<b>BS=64</b>	<b>BS=32</b>	<b>BS=16</b>	<b>BS=8</b>	<b>BS=4</b>	<b>BS=2</b>
<b>LR = 5e-4</b>	1.38	1.38	1.38	1.35	1.36	1.36
<b>LR = 1e-3</b>	1.40	1.37	1.36	1.22	1.18	1.21
<b>LR = 5e-3</b>	1.10	0.99	0.90	0.95	1.00	1.16
<b>LR = 1e-2</b>	0.88	0.83	0.85	0.90	1.03	1.17
<b>LR = 5e-2</b>	0.73	0.75	0.84	0.90	1.02	1.15

**Table 4.9: RMSE result of each batch size and learning rate with weight decay value 10**

<b>RMSE Table</b>	<b>BS=64</b>	<b>BS=32</b>	<b>BS=16</b>	<b>BS=8</b>	<b>BS=4</b>	<b>BS=2</b>
<b>LR = 5e-4</b>	1.38	1.36	1.39	1.35	1.35	1.39
<b>LR = 1e-3</b>	1.39	1.43	1.37	1.37	1.35	1.41
<b>LR = 5e-3</b>	1.33	1.36	1.34	1.31	1.36	1.39
<b>LR = 1e-2</b>	1.34	1.33	1.31	1.36	1.36	1.40
<b>LR = 5e-2</b>	1.33	1.32	1.34	1.35	1.43	1.42

In order to test the impact of the security awareness level on the result, the model has been trained with and without the user's security awareness level and score as a feature. Each model has been trained 10 times with 10 epochs. Table 4.10 shows the RMSE results

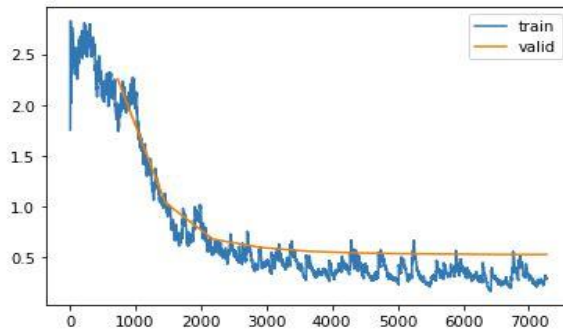


of 10 times training. The sample of model training results with and without feature is shown in Figure 4.1 and Figure 4.2.

**Table 4.10: RMSE result of ISA-ware model and Non-ISA-aware model.**

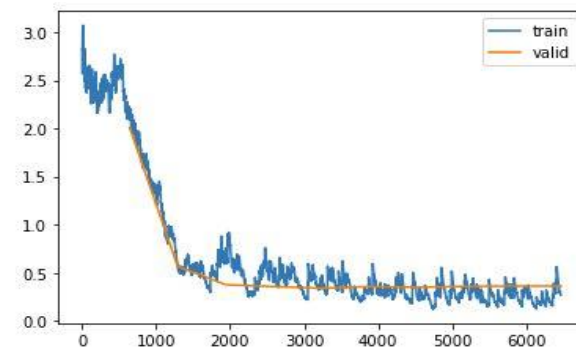
	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>	5 <sup>th</sup>	6 <sup>th</sup>	7 <sup>th</sup>	8 <sup>th</sup>	9 <sup>th</sup>	10 <sup>th</sup>
With feature	0.73	0.74	0.73	0.73	0.72	0.75	0.73	0.74	0.73	0.73
Without feature	0.61	0.61	0.60	0.61	0.62	0.60	0.61	0.61	0.60	0.61

epoch	train_loss	valid_loss	_rmse	time
0	0.477407	0.521804	0.722360	00:05
1	0.308583	0.520179	0.721234	00:05
2	0.312033	0.519253	0.720592	00:05
3	0.198531	0.531443	0.729001	00:05
4	0.565255	0.513904	0.716871	00:05
5	0.218809	0.527326	0.726173	00:05
6	0.160855	0.533450	0.730376	00:05
7	0.183894	0.536371	0.732373	00:05
8	0.163644	0.540433	0.735141	00:05
9	0.103151	0.540890	0.735452	00:05



**Figure 0.1: Sample of training result of without ISA**

epoch	train_loss	valid_loss	_rmse	time
0	2.097552	2.005742	1.416242	00:06
1	0.581065	0.567288	0.753185	00:06
2	0.754696	0.377205	0.614170	00:06
3	0.422112	0.357324	0.597766	00:06
4	0.380697	0.340268	0.583325	00:06
5	0.417702	0.357510	0.597922	00:06
6	0.262937	0.349617	0.591284	00:06
7	0.385875	0.359568	0.599640	00:06
8	0.273463	0.362088	0.601738	00:06
9	0.269888	0.362727	0.602268	00:06



**Figure 0.2: Sample of training result with ISA**

From the result, it is find that the security awareness level and score are important features to the recommendation system as the average RMSE without the security awareness score as a feature is 0.73 while the average RMSE with the security awareness

score as a feature is 0.61. The RMSE result improved significantly after use security awareness level and score as features.

The accuracy of the recommendation system is determined using a confusion matrix. The confusion matrix is generated by using 20% of the dataset as a validation dataset to test the model. There are 10 times of testing to get the average value for confusion matrix. The value of the predicted rating is decimal instead of an integer, the predicted rating has to be rounded before generate the confusion matrix. The ratings with value in range 1 to 3 are considered as not preferred or not recommended, rating value in range 4 to 5 is considered as preferred or recommended. For example, the recommendation system gives a predicted rating for user U1 on recommendation R1, the value of prediction is 3.5964. The predicted rating needs to be rounded before generate the confusion matrix. In this case, the value will be 4, and the rating value with 4 is considered as ‘Recommended’. The confusion matrix of the model with and without using security awareness as features are shown in Table 4.11 and Table 4.12. Table 4.13 shows the confusion matrix on second round of profiling.

**Table 4.11: Confusion matrix of the model without features**

	<b>Recommended</b>	<b>Not recommended</b>
<b>Preferred</b>	85 (True-Positive)	21 (False-Negative)
<b>Not preferred</b>	8 (False-Positive)	290 (True-Negative)

**Table 4.12: Confusion matrix of the model with features**

	<b>Recommended</b>	<b>Not recommended</b>
<b>Preferred</b>	97 (True-Positive)	13 (False-Negative)
<b>Not preferred</b>	2 (False-Positive)	292 (True-Negative)

**Table 4.13: Confusion matrix of the model with feature on second profiling**

	<b>Recommended</b>	<b>Not recommended</b>
<b>Preferred</b>	94 (True-Positive)	9 (False-Negative)
<b>Not preferred</b>	2 (False-Positive)	294 (True-Negative)

The accuracy of the model without security awareness is 85.72% while the accuracy of the model with security awareness is 97.92%. The security awareness helps the recommendation system increase 12.2% of accuracy. Moreover, the accuracy is further slightly improved on second round of profiling by 0.74%.

#### **4.4 Implications**

From the result, it is find that people with similar security awareness score and level are having similar preferences for recommendations. Each of the triggered recommendations has an average rating higher than 3 which meaning that users prefer those recommendations which point out the issue of the computer clearly, and addressed the problem of ambiguous recommendations. This describes the importance to have a recommendation system in the cybersecurity domain for the public in Malaysia to improve security awareness.

With the security awareness profile, the accuracy of the recommendation system improved significantly. This explains 2 important results. Firstly, determining security awareness using the measurement method without a questionnaire is feasible as it successfully classifies the people with security awareness levels. The ISA profiling system is able to collect necessary data from the users automatically without questionnaire, which helps to resolve the problem of Hawthorne effect. In addition, the structure of the ISA profiling system supports ISA measurement for public as the profiling process can be implemented through Internet. It expanded the scope of application from an organization to public while reduced the time of measurement.

Secondly, profiling users using security awareness can help the collaborative filtering recommendation system to provide a more accurate prediction.

#### **4.5 Summary**

In this chapter, the statistical analysis result of data was presented. The result of RMSE and confusion matrix were presented and analyzed. The implications concluded that the security awareness measurement without a questionnaire is feasible and profiling users using security awareness can help in the recommendation system. The profiling system slightly increase the users ISA result. The model with ISA-aware has a better RMSE and accuracy compare with the model without ISA-aware, the RMSE is improved 0.12 and accuracy 12.2%.

## **CHAPTER 5 CONCLUSION**

### **5.1 Overview**

This chapter summarizes and concludes the result of this research. The contributions and suggested future works are presented in this chapter. The problems of ambiguous recommendation, complexed recommendation, lack of ISA measurement for individual, and measurement that cause Hawthorne effect are addressed by the ISA profiling system and collaborative recommendation system that proposed in this research.

### **5.2 Results**

The results of the experiment were determined by RMSE and the accuracy of the model. The RMSE and accuracy result of the model with security awareness is better than the model without security awareness. Profiling user using security awareness is able to find out the user's recommendation preference.

### **5.3 Contribution**

The contribution of this research was a framework of profile users using security awareness and create a collaborative recommendation system to provide a suitable recommendation to the users. The framework created in this research can be expanded for future work. This research shows the relationship between security awareness and

recommendation preference. A collaborative recommendation system could be created to increase the security awareness of Malaysians.

#### **5.4 Future Work**

The process of data collection is not fully automated. Users are required to enter the email address, password, and social media account. A full automation data collection process needs to be explored to increase the willingness to use. Also, this framework presented a basic measure on each focus area, more measures could be collected from each focus area.

The recommendation system has the potential to be a very useful tool to increase the security awareness of the public. Instead of using only collaborative filtering to provide recommendations, a hybrid recommendation system could be implemented. Instead of using a rule-based technique, a content-based or knowledge-based recommendation system to utilize computer health to points out the vulnerability to address the problem of providing the ambiguous recommendation should be built.

A real-time architecture should be implemented on the recommendation system. Immediate response can provide more insightful recommendations to users. More recommendations should be added to the list of the recommendation system. Since the users have been classified using security awareness, recommendations with different depths should be considered to be added. A better recommendation system can be developed with more refinement.

#### **5.5 Summary**

This research has shown that security awareness can be used for user profiling and collaborative recommendation systems. The results show that people with similar security awareness are having a similar preference for recommendations. The results also show that recommendations can be provided to users accurately using collaborative filtering and the security awareness profile. There are many refinements to develop a better framework in this area of study.

## REFERENCES

- Ariffin, M. R. K., & Letchumanan, M. (2020). Status of Cybersecurity Awareness Level in Malaysia. In *Innovations in Cybersecurity Education*. [https://doi.org/10.1007/978-3-030-50244-7\\_17](https://doi.org/10.1007/978-3-030-50244-7_17)
- Bada, M., Sasse, A. M., & Nurse, J. R. C. (2019). Cyber Security Awareness Campaigns: Why do they fail to change behaviour? In *arXiv*.
- Bendovschi, A. (2015). Cyber-Attacks – Trends, Patterns and Security Countermeasures. *Procedia Economics and Finance*. [https://doi.org/10.1016/s2212-5671\(15\)01077-1](https://doi.org/10.1016/s2212-5671(15)01077-1)
- Cyber Security Awareness Alliance. (n.d.). *5 Simple Ways You Can Fight Spam and Protect Yourself*. Cyber Security Awareness Alliance. Retrieved April 27, 2021, from <https://www.csa.gov.sg/gosafeonline/go-safe-for-me/homeinternetusers/5-simple-ways-you-can-fight-spam-and-protect-yourself>
- Dashika Gnaneswaran. (2018). *Cybersecurity threats to cost organizations in Malaysia US\$12.2 billion in economic losses*. Microsoft Malaysia. <https://news.microsoft.com/en-my/2018/07/12/cybersecurity-threats-to-cost-organizations-in-malaysia-us12-2-billion-in-economic-losses/#:~:text=The study reveals that the,GDP of US%24296 billion><sup>1</sup>.
- Donovan, K. (2017). *10 Cybersecurity Best Practices for IT, IS, Network, Data Security | ObserveIT*. April 25.
- DSP Mahfuz Bin Dato' Ab. Majid. (2013). Cybercrime: Malaysia. *Royal Malaysia Police*, 19. <http://www.skmm.gov.my/skmmgovmy/media/General/pdf/DSP-Mahfuz-Majid-Cybercrime-Malaysia.pdf>
- Fernando, S. A., & Yukawa, T. (2014). Securing Information Sharing Through User Security Behavioral Profiling. In *Transactions on Engineering Technologies*. [https://doi.org/10.1007/978-94-017-8832-8\\_47](https://doi.org/10.1007/978-94-017-8832-8_47)
- Frey, B. B. (2018). Confirmatory Factor Analysis. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, Istmet 2014*, 218–223. <https://doi.org/10.4135/9781506326139.n140>

- Galba, T., Solic, K., & Lukic, I. (2015). An information security and privacy self-assessment (ISPSA) tool for internet users. *Acta Polytechnica Hungarica*. <https://doi.org/10.12700/aph.12.7.2015.7.9>
- Hart, S., Margheri, A., Paci, F., & Sassone, V. (2020). Riskio: A Serious Game for Cyber Security Awareness and Education. *Computers and Security*. <https://doi.org/10.1016/j.cose.2020.101827>
- Juno Risk Solutions. (2015). *Cybersecurity Best Practices Guide For IIROC Dealer Members*. 1–53.
- Khan, A. H., Sawhney, P. B., Das, S., & Pandey, D. (2020). SartCyber Security Awareness Measurement Model (APAT). *2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and Its Control, PARC 2020*. <https://doi.org/10.1109/PARC49193.2020.236614>
- Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Protecting people from phishing: The design and evaluation of an embedded training email system. *Conference on Human Factors in Computing Systems - Proceedings, June 2014*, 905–914. <https://doi.org/10.1145/1240624.1240760>
- Li, H., Cai, F., & Liao, Z. (2019). Content-Based Filtering Recommendation Algorithm Using HMM. *2012 Fourth International Conference on Computational and Information Sciences, March*, 275–277. <https://doi.org/10.1109/ICCIS.2012.112>
- Litoussi, M., Kannouf, N., El Makkaoui, K., Ezzati, A., & Fartitchou, M. (2020). IoT security: challenges and countermeasures. *Procedia Computer Science*, *177*, 503–508. <https://doi.org/https://doi.org/10.1016/j.procs.2020.10.069>
- Lorenzi, F., & Ricci, F. (2003). *Case-Based Recommender Systems : A Unifying View*. *Case-Based Recommender Systems : a Unifying View. September 2014*. <https://doi.org/10.1007/11577935>
- Lyons, K. B. (2014). *A Recommender System In The Cyber Defense Domain*. 81. <https://doi.org/10.1002/ajmg.b.31008>
- McCarthy, C., Harnett, K., & Carter, A. (2014). A Summary of Cybersecurity Best Practices. In *NHTSA*.

- Mui, L. Y., Aziz, A. R. A., Ni, A. C., Yee, W. C., & Lay, W. S. (2002). A survey of Internet usage in the Malaysian construction industry. In *Electronic Journal of Information Technology in Construction* (Vol. 7, Issue December, pp. 259–269).
- Muniandy, L., Muniandy, B., & Samsudin, Z. (2017). Cyber Security Behaviour among Higher Education Students in Malaysia. *Journal of Information Assurance & Cybersecurity, February 2017*, 1–13. <https://doi.org/10.5171/2017.800299>
- Naidoo, R. (2020). A multi-level influence model of COVID-19 themed cybercrime. *European Journal of Information Systems*. <https://doi.org/10.1080/0960085X.2020.1771222>
- Ngoqo, B., & Flowerday, S. V. (2015). Information Security Behaviour Profiling Framework (ISBPF) for student mobile phone users. *Computers and Security*. <https://doi.org/10.1016/j.cose.2015.05.011>
- Parsons, K., McCormac, A., Butavicius, M., Pattinson, M., & Jerram, C. (2014). Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q). *Computers and Security*. <https://doi.org/10.1016/j.cose.2013.12.003>
- Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2014). A study of information security awareness in Australian government organisations. *Information Management and Computer Security*. <https://doi.org/10.1108/IMCS-10-2013-0078>
- Rahman, A., Lubis, M., & Ridho, A. (2015). Information Security Awareness at the Knowledge-Based Institution : Its Antecedents and Measures. *Procedia - Procedia Computer Science*, 72, 361–373. <https://doi.org/10.1016/j.procs.2015.12.151>
- Reep-van den Bergh, C. M. M., & Junger, M. (2018). Victims of cybercrime in Europe: a review of victim surveys. *Crime Science*. <https://doi.org/10.1186/s40163-018-0079-3>
- Ruzgar, N. S. (2005). A research on the purpose of internet usage and learning via internet. *The Turkish Online Journal of Educational Technology*, 4(4), 27–32. <https://www.mendeley.com/catalogue/04240f58-c59f-32ed-81e3-6339542c806a/>
- Salton, G., Wong, A., & Yang, C. S. (1975). *A Vector Space Model for Automatic Indexing*.



18(11).

- Sawaneh, I. A. (2020). Cybercrimes: Threats, Challenges, Awareness, and Solutions in Sierra Leone. *Asian Journal of Interdisciplinary Research*. <https://doi.org/10.34256/ajir20114>
- Shaw, R. S., Chen, C. C., Harris, A. L., & Huang, H. J. (2009). The impact of information richness on information security awareness training effectiveness. *Computers and Education*. <https://doi.org/10.1016/j.compedu.2008.06.011>
- Tan, O., Leng, S., Vergara, R. G., Khan, N., & Khan, S. (2020). Cybersecurity and Privacy Impact on Older Persons Amid COVID-19: A Socio-Legal Study in Malaysia. *Asian Journal of Research in Education and Social Sciences*.
- Tharshini, N. K., Hassan, Z., & Mas'ud, F. H. (2021). Cybercrime threat landscape amid the movement control order in Malaysia. *International Journal of Business and Society*, 22(3), 1589–1601. <https://doi.org/10.33736/ijbs.4323.2021>
- Velki, T., Solic, K., & Ocevcic, H. (2014a). Development of Users' Information Security Awareness Questionnaire (UISAQ) - Ongoing work. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings*. <https://doi.org/10.1109/MIPRO.2014.6859789>
- Velki, T., Solic, K., & Ocevcic, H. (2014b). Development of Users' Information Security Awareness Questionnaire (UISAQ) - Ongoing work. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings, May*, 1417–1421. <https://doi.org/10.1109/MIPRO.2014.6859789>
- Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *Computers and Security*. <https://doi.org/10.1016/j.cose.2013.04.004>
- Wahyudiwan, D. D. H., Sucahyo, Y. G., & Gandhi, A. (2017). Information security awareness level measurement for employee: Case study at ministry of research, technology, and higher education. *Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education*,

*Industry and Society in Big Data Era, ICSITech 2017, 2018-Janua*, 654–658.  
<https://doi.org/10.1109/ICSITech.2017.8257194>

WaterISAC. (2015). *10 Basic Cybersecurity Measures: Best Practices To Reduce Exploitable Weaknesses And Attacks*. June, 9. [https://ics-cert.us-cert.gov/sites/default/files/documents/10\\_Basic\\_Cybersecurity\\_Measures-WaterISAC\\_June2015\\_S508C.pdf](https://ics-cert.us-cert.gov/sites/default/files/documents/10_Basic_Cybersecurity_Measures-WaterISAC_June2015_S508C.pdf)

Yao, J., Hauptmann, A. G., & Post, W. (2018). *News Recommendation and Filter Bubble Politico*. 1–4.

Zanker, M. (2010). *Introduction to Recommender Systems About the speakers*. March, 1–139.

Zhang-Kennedy, L., & Chiasson, S. (2021). A Systematic Review of Multimedia Tools for Cybersecurity Awareness and Education. In *ACM Computing Surveys*.  
<https://doi.org/10.1145/3427920>

Zwilling, M., Klien, G., Lesjak, D., Wiechetek, Ł., Cetin, F., & Basim, H. N. (2020). Cyber Security Awareness, Knowledge and Behavior: A Comparative Study. *Journal of Computer Information Systems*. <https://doi.org/10.1080/08874417.2020.1712269>