CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter purposely is to study, review and identify all the related theories and facts required to be applied in this research. Designing, developing and implementing a risk assessment for text spam message required certain scopes of theories that include issues related to spam threat, Danger Theory (DT) of Artificial Immune System (AIS), risk management, text mining and dataset deployment during the testing phase.

Prior to design and develop the risk assessment of text spam messages, every topic that is closely related and required for the research is discussed. The review is conducted by gathering research papers and journals from numerous and reliable sources for the subject as aforementioned. Besides reviewing the existing works, this session also tries to uncover the gap between the past and current situation. Through this, a potential solution for the identified problem and gap is unfolded, then suggested in this research. At the final section of this chapter, the taxonomy of the research direction is elaborated. This is important and it becomes the guidance to execute this research by phases.

2.2 The Issue Of Spam

2.2.1 Spam History

A junk message, colloquially known as spam, has been interfering human daily life since the year of 1990's. Since then, this threat has become as one of the significant affair globally. At an early stage, it appeared as a harmless message but annoying such as sale advertisement in electronic mail or e-mail form. However, this threat becoming vicious since these irritating messages has evolved into a criminal landscape.

According to Internet Society (n.a., 2014b), a common definition of spam is unsolicited bulk messages, that is, messages sent to multiple recipients who did not ask for them. Natris (2014) from Internet Governance Forum or IGF elaborates more that unsolicited bulk messages have different meaning depending on the country.

However, this bulk SMS is not appeared to be as spam for certain circumstances (Knipe, 2017). For instance, the emergency services of the Australian government is using bulk SMS as the most effective digital communication medium to inform users in the high fire risk area about latest update such as evacuate plan in aggressive fire season. While in the United States, bulk SMS are sent out to alert users, within a targeted area, of an Amber alert or a matter of public emergency.

The approach taken in spam handling by these countries are also varies. It is also defined as unsolicited electronic communications because of its nature occurred with the assistance of electronic platforms such as the Internet and mobile technology.

The history of spam is closely tied to the history and evolution of the Internet itself. This so-called co-evolutionary threat also becoming more vicious and expand when there is a success achievement in the integration of the Internet and mobile technology.

Looking back in 1993, the first time use of the term spam was for a post from USENET by Richard Depew (n.a., 2014a). This was the resulted of a bug in a software program that caused 200 messages to go out to the newsgroup. In 1997, the very first Simple Mail Transfer Protocol or SMTP has been hijacked

where this is a sender push technology that delivered all messages without any sender requirement to provide authentic return addresses. This has made the job of spamming much easier for the spammer. This spam in e-mail form has gone worse in 2001 when Code Red worm and Sircam virus infiltrate thousands of web servers and e-mail accounts causing a spike in Internet bandwidth usage. Two years later, it was the first time that the amount of spam e-mail exceeded the amount of legitimate email. The first spamming "botnets" also appeared in this year of 2003. Researchers found that most of the spam sent around the world was in English, however, spammers started using automatic translations services to send spam in other languages in 2009. Then, in 2012 there is a rising number of social media spam, parallel with the advancement of Internet and mobile technology integration.

In a short while, the rise of spam has evolved into something that is unprecedented such as instant messenger services, fax to email services, Voice over Internet Protocol (VoIP), mobile and smartphones, social networks and mobile instant messenger applications (Natris, 2014). Spammers are easily adapt to use the available technology to reap the potential illicit revenue. Even though there are many mechanisms has been applied to curb this threat, its adverse effect is still kept rising persistently, with no sign to be any lesser.

2.2.2 Spam Characteristics And Its Adverse Effects

Basically, spam has unique characters that can be recognized regardless of its format; either in an e-mail, SMS, or social media spam. In 2014, a group of experts from the Internet Governance Forum or IGF consent that the message or communication perceived as spam if they are possessed of this essential quality (Natris, 2014).

i. nuisance;
ii. considered as an invasion of privacy;
iii. considered offensive;
iv. contain embedded malware or spyware;
v. aims to mislead or deceive, with the potential to cause financial loss, identity theft, and cause other harm;

vi. may inflict direct financial costs; for instance, where internet access or replying SMS message is charged.

In addition to this, China via East West Institute and Internet Society of China (Rauscher & Yonglin, 2011) describe spam as being uninvited by the recipient, high in volume and distributed widely; regardless of electronic form which includes email, instant messaging, web search engine, fax, Internet site postings, mobile texting, SMS, and tweeting as well as others.

This common feature of spam should be easy to be differentiated from non-spam or ham messages; however, statistic showed that spam is not easy to fend off. Its adverse effects have been recorded and continually rising from around the globe. It has been noticed as one of the security threats whereby it used to be a nuisance message but currently has become more aggressive since it now acts as a potential criminal toolkit. This claim is supported by a comprehensive remark from IGF (Natris, 2014), that the content and intent of spam in 2014 is more violent, intrusive and malicious than in 2000 which is spam is now used for all sorts of intrusions, malware spreading, fraudulent messages and phishing attacks.

The wide proliferation and advancement of mobile devices have attracted the attention of cyber-criminals, who exploited the functionality of the device for malevolent purposes. In 2016, an annual threat report by Cloudmark showed an alarming situation in the United States of America (USA) and United Kingdom (UK) which already expand its effects throughout the world. These findings include SMS and email spam as the following:

i. phishing attempts accounted for 20% of SMS spam reported in the USA;
ii. 43% of all SMS spam seen in the UK pushed payday loan offers;
iii. 67% of SMS spam messages lured USA victims with various financial offers;
iv. 85% of SMS reports in the UK used money as their pitch;
v. spammers had compromised well over 200,000 websites over the past year; and
vi. the USA generates about a third of the world's email spam.

The negative impact of spam can be affected by many aspects such as money loss, discredit of reputation, and even time wasting. These adverse effects are borne by individuals, organizations and also governments (Natris, 2014). On top of that, a study of assessing possible risk impact on using smartphone has been done by Alotaibi, Furnell, & Clarke (2017) identified more effects other than financial loss such as embarrassment, breach of personal privacy, breach of commercial confidentiality, legal liability, threat of personal safety and disruption of services.

For instance, spam able to risks in diminishing trust in doing business online, which has a direct effect to the economy. Potential clients that have doubt or untrustworthy sense towards online business would contribute to the partial cause of an economic slowdown. In other situation, the Malaysian Communications and Multimedia Commission (MCMC) revealed that many users are unaware that they were being charged for every SMS spam message in between 50 cents to RM3 (n.a., 2015).

However, information loss or damage is not always as obvious or observable as a financial loss (Zhang et al., 2011). End users are directly affected because of the time spent for dealing with spam messages. Prior to identifying spam, they need to read through and revise the entire received message so they will not overlook any legitimate messages that possibly important to them. In addition to that, the cost that they need to spent on protective measures against spam and also potentially the loss through fraudulent actions that caused from responding to the deceitful message. There was a case reported by a local newspaper in Malaysia (n.a., 2016) where an individual that received an SMS message informing that he/she required calling the bank for a verification of credit card transaction. Unfortunately, he/she gets deceived and the call was answered by a scammer. With the convincing method of social engineering technique, the victim declares all the sensitive information without realizing that he/she already lose the money.

While for the government, there is a financial impact on the implementation of consumer awareness campaigns, legislative processes, regulatory and enforcement agencies. CyberSecurity Malaysia[1] reported (n.a.,

[1] www.cybersecurity.my

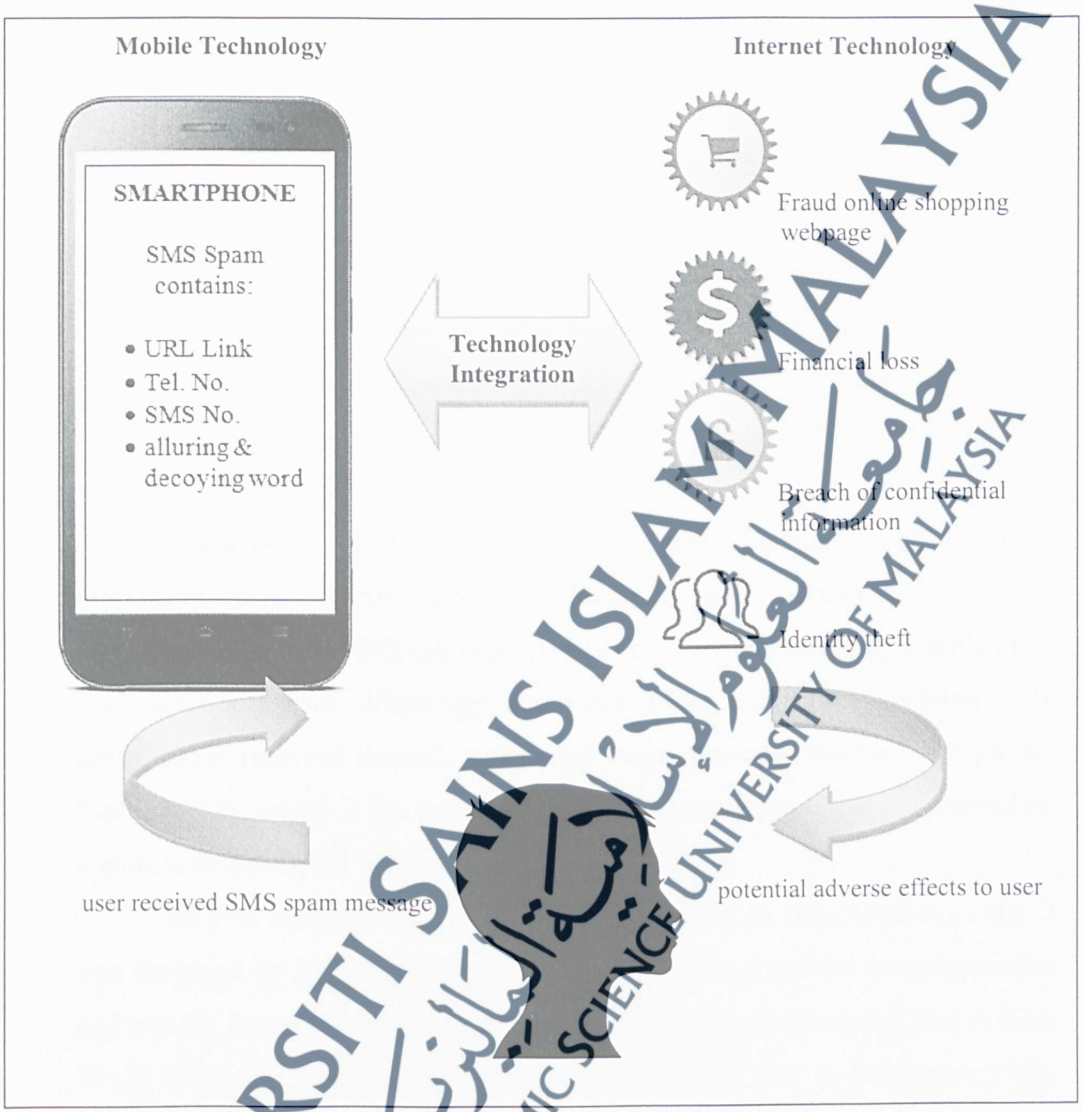
2017) that there is a tremendous increment in the case of online fraud in between 2015 to 2016 which include online business, online banking transaction, phishing, and scam.

The Internet and mobile industries are also affected somehow in several ways. The cost impact of spam is significant for email service providers, which require increased storage capacity, faster processing capability, and access to higher levels of bandwidth to manage the ever-increasing volume of traffic. Internet services, hosting and platform providers invest in different protective measures. There is a need for the whole industry to works together to develop Internet standards, best practices, and codes of conduct for their use. Telecommunication service providers also are not being left behind in providing the cost of preventive measurements such as the installation of the anti-spam solution or investing in any reliable filtering techniques (Abdulhamid et al., 2017), to minimized the amount of spam.

The advancement of mobile devices such as smartphone that is integrated with Internet technology has made the usefulness of SMS has become wider. The link provided in SMS is easily clickable and accessible online using smartphone. Through this facility, users are prone to cyber crime activities. The definition of smartphone as clarified in Theoharidou, Mylonas, & Gritzalis (2016) is a mobile phone with advanced capabilities which executes an identifiable operating system allowing users to extend its functionality with third party applications that are available from an application repository. Besides spam attacks, this paper elaborated another type of attacks that could risk smartphone users which include eavesdropping, unauthorized access and disclosure of sensitive information.

There is a well-known method in assisting spammer to make their spam messages become even more successful, namely as SMiShing. Yeboah-Boateng & Amanor (2014) articulated that SMiShing takes place whenever text messages are sent to the user to either click on a link provided which leads to a fraudulent website or also can allow the attacker to get access to user's confidential information. Originally this attack kind of attack is known as phishing which is occurred through emails. SMiShing becomes visible in the form of phishing that uses SMS or text messages on smartphone (Jain & Gupta, 2018).

Moreover, the proliferation of mobile technology has stimulated telecommunication provider in offering a cheaper cost for Internet access. In addition to that, at these days wireless networking technology of Wi-Fi facility is accessible almost everywhere with at no cost at all.



Source: Natris (2014); Yeboah-Boateng & Amanor (2014); Zhang et al. (2011)

Figure 2.1: The Integration Of Mobile (Smartphone) And Internet Technology

Spam also has carried vicious attack to scare victims such as disseminating ransomware in order to achieve what these ransomware criminals want. According to report Scott & Spaniel (2016), this type of attack is so effective is that the cyber security field is not entirely prepared for its

resurgence. Previously shown that spam email has become as it one of delivery channel and the trends also anticipated that SMS spam could be as it major transport in this coming year.

2.2.3 Why Focus On SMS Spam

A different type of spam that frequently encountered today is discussed in Vural & Venter (2012). Authors of this paper described comprehensively that there are a few type of spam at present, which are email, comment, messaging, mobile and VoIP spam. Email spam is well known as the initial version of spam and the most common in the computer world. The fast development of Internet technology that leads to the emergence of web pages has caused comment spam to arise. This kind of spam usually inflicts the comments section of newspaper websites, where adverts are inserted in the comments section. Then, advancement of mobile technology and emergence of smartphone has made messaging spam become increased. Messaging spam is also as SPam over Instant Messaging (SPIM) that is received over an instant messenger application such as Google Talk, WhatsApp, and Viber. Later, users start receiving VoIP spam that is received through automated voice messages over a VoIP phone. Mobile spam, which is the essence of this research, is spam that is received on a mobile device in the form of SMS messages.

SMS is text messages with limited content up to 160 characters only. It was designed by Hilderbrandt to accommodate a short mobile communication and was the first messaging platform in the era of communication (Chen & Kan, 2013). Since the first ever mobile text message was sent in December 1992, SMS has witnessed phenomenal growth, and around the world, and now seen that more 7 trillion SMS messages exchanged every year. In just 20 years, SMS started to become the most widely used form of written electronic communication (Lota & Hossain, 2017; PortioResearch, 2015). Nowadays, SMS can be used as a platform for user interaction mode in many applications such as password verification, voting, SMS remote control, SMS alert and banking (Rahman, Abdullah, A. Hamid, Wen, & Mohd Jelani, 2017; Singhal, Arora, Kumari, & Majumder, 2013).

The availability of unlimited prepaid SMS packages with cheap cost has made it as an easy target for spammers to disseminate spam messages. In addition to that, SMS is considered as a trusted service with subscribers comfortable using it for confidential information exchange. As a result, SMS can result in a higher response rate for spam than email (Kim, Jo, & Choi, 2015; Delany, Buckley, & Greene, 2012).

A survey done by Mobile Ecosystem Forum or MEF[2] found that SMS is strongly active and required by most users as their messaging platform. This survey is a study of messaging behavior that covers nine (9) countries; United State of America, United Kingdom, Brazil, France, Germany, China, India, South Africa and Nigeria. From this study, it is revealed that SMS is still a usable and users likely to depend upon as a messaging platform (Malcolm, 2016).

i. SMS outscores messaging apps for Application to Person or 'A2P';
ii. financial services lead the way for contact by SMS;
iii. most popular SMS use for business purposes is confirming a password; and
iv. SMS is the mostly utilized for the enterprise to consumer purposes.

Although this survey observed that voice and face-to-face conversation still popular channels for person-to-business communications, SMS become the favorite option when the need for physical appearance is not required at all.

Even though there is an emergence of many other mobile messaging platforms such as WhatsApp and WeChat, a study for a pattern of SMS usage that is executed by GSMA[3] surprisingly anticipated that there is increasing rate for SMS usage in the future, together with the sophistication of spam attack globally (Cloudmark, 2011). Worsening the situation, high usage, and dependant of SMS platform for text messaging will widen the adverse effects, adjacently with the proliferation of mobile and Internet technology. This claim is also supported by

[2] MEF is a global trade body, established in 2000, acts as an impartial and authoritative champion for addressing issues affecting the broadening mobile ecosystem. The goal is to accelerate the growth of a sustainable mobile ecosystem that drives inclusion for all and delivers trusted services that enrich the lives of consumers worldwide. http://www.mobileecosystemforum.com

[3] The GSMA represents the interests of mobile operators worldwide, uniting nearly 800 operators with almost 300 companies in the broader mobile ecosystem, including handset and device makers, software companies, equipment providers and internet companies, as well as organizations in adjacent industry sectors. http://www.gsma.com

a survey done by Portio Research Limited[4], which 6.1 billion users, out of a total human population of 7.3 billion worldwide, or 84% use SMS as a communication platform in 2015. This number is still rising and will peak at around 6.4 to 6.5 billion over the next three (3) years. SMS will remain a major standard for more than 6 billion users for most of the next decade (PortioResearch, 2015). The rising numbers of SMS messages are likely to boost an increasing amount in the development of spam messages also. The same number of high usage of SMS platform as the communication medium also has been reported in Ezpeleta, Garitano, Zurutuza, & Hidalgo (2017).

Chen & Kan (2013) pointed out that the SMS and Twitter usage globally makes 37.83% and 0.16% respectively. But the research to study SMS only accounts for 14.29%, while 75% for Twitter. This statistic somehow showed that there are still lacking figure in studying SMS even though the utilization of the services is high.

In addition to that, as reported by Neil Cook who is the Chief Technology Officer of security firm, Cloudmark, suggested that SMS malware could represent a much more pressing danger than malicious emails. This claim is based on findings that smartphone are currently a more trusted medium than email, making users more susceptible to phishing scams (n.a., 2013).

2.3 Human Perception And Behaviour Towards Online Threat

In many studies, human is the factor that is identified as one of the vulnerabilities in managing information security. Even in most studies, researchers agreed that human is the weakest link in this situation. A study was done by Yeboah-Boateng & Amanor (2014) somehow has resulted in supporting this claim. As reported by Lin et al. (2012), this is due to reason that users being more interested and trusting in web sites based on its visual appearance and lack of knowledge on the various security features are the most common weakness that easily gets exploited by spammers. Some more evidence, a survey done by Fallows (2011) found that trust is the backbone to make the most

[4] Portio Research Ltd. Is an independent telecoms research provider that is specialize in mobile messaging. http://www.portioresearch.com

transaction via the Internet and Internet users commonly look for web sites that look convincing to them. A well-known hacker that currently is a corporate security consultant has the same claim in his book, 'The Art of Deception: Controlling the Human Element of Security'. He strongly affirmed that human is the weakest element in security; mostly caused by misguided trust and therefore leading many users to settle for a false sense of security (Mitnick & Simon, 2003). Furthermore, a study by Goel & Jain (2017) shown that 35% of users choose SMS as the most trusted messaging platform, followed by 28% by messaging applications such as Whatsapp, 18% by Facebook, Yahoo Messenger and Skype, and 16% is from push notification such as for verification of bank transaction. A study by Haritha, Kumar, & Krishnan (2017) showed that 11.1% of the spam's recipients had responded to the SMS spam message. Although this response rate is numerically low, unpredictably the amount of money loss has caused billions of dollars.

A report from Chief Information Security Officer (CISO) workshop (Johnson & Moag, 2011) declared that 80% of the respondents agreed that the human-related risks are more troublesome than the technical challenges. Human factor threats are pervasive at all levels and it is inevitable. For instance, Jayakumar & Phippen (2006) described that the perception of online issues differs from each and every individual like normal user, Internet user and security professional because the normal user just uses the system without having much knowledge about the threats and causes whereas the Internet users are known to some available threats like virus, spyware, and malware. But the security professionals are known to all kinds of threats that are available and its causes.

Today and certainly in the future, users-related risks are considered to be the greatest risk at all levels within the organization. These types of risk could be perpetrated by users who using IT facilities, deploying operational facilities and processes, or by users who having privileges or management responsibility (Humphreys, 2010).

To teach and educate workers in organizations or any individuals to recognize risk is a large part of the challenge in exposing them the awareness towards online threats. The conclusion from a workshop is that the best defence is to have a good offense such as educate the users and manage the data wisely (Johnson & Moag, 2011). Authors for paper 'Strengthening the Human Firewall' (Tjhai & Furnell, 2006) insisted that complacency, ignorance, and unawareness of security are amongst the biggest

obstacles to maintain IT security within an organization, which include in the domain of individuals security. Indeed, technical security controls alone are not enough to provide a real protection if there is no human participation acquired in this stage. As a consequence, it is desirable for many organizations to develop an effective security awareness program; by which user awareness could be enhanced. The survey done showed that 70% of respondents do not really know how social engineering is working on them, only 8% had firmly adhered to the password policy by applying all strong password characteristics, and surprisingly, two-thirds of respondents considered that the IT department has the sole responsibility for securing corporate IT systems.

In 2010, Securelist produced a report (David Emm, 2010) that is related to online threats, which is suggesting that human vulnerabilities need to be patched seriously. Users are not only susceptible due to a lack of awareness, but sometimes they get lured by the content of spam messages and able to entice any of them into clicking on a link that should be ignored. Common sense often suggests that if something seems too good to be true, it probably is. However, the same common sense may not result in the understanding that taking action, in this case, clicking on a link could be harmful.

A survey done by Zhang et al. (2011) concluded that user awareness is the best defence in managing and handling cyber-crimes. This also supported by a study done by Yeboah-Boateng & Amanor (2014) which discovered that users usually perceived as unwary or oblivious of the numerous cyber-attacks against their mobile devices. This study also disclosed that most users are either slightly or not concerned at all with cyber-attacks, which is surprisingly 65% of the respondents in this survey felt comfortable in providing their personal credentials online. The same findings that suggesting human is easily influenced by cyber-attacks is a research done by Lin et al. (2012). Authors studied the user's perceptions of whether a given action is legitimate, or how that action makes them feel with respect to privacy. Educate users in terms of security awareness and training is also strongly suggested by Colwill (2009) and considered as a critical mechanism to be implemented in securing the confidential data and information.

Bujang & Hussin (2012) suggested on their study to understand user's behaviour and ethics perspectives towards email spam prior developing any mechanism of the solution. With the application of Technology Threat Avoidance Theory (TTAT), they are proposing that effectiveness of spam control measures can be improved based on an understanding of how the users perceive spam as a threat. Awareness of lack of

knowledge or skill in the certain application will make them more cautious in the technology.

As elaborated in Colwill (2009), the application of technology alone will not provide solutions. Security controls need to be workable in a variety of environments and designed, implemented and maintained with users' behaviour in mind.

2.3.1 The Necessity Of Users' Assistance

A survey has been conducted to identify the necessity of assistance for users to detect malicious messages. Three (3) high-levels professionals from government sectors has participated in this survey with the objective to identify the level of demand of the proposed model in this study, Risk Concentration for Context Assessment (RiCCA) in daily life of mobile users. From this survey, the implementation of the proposed model is potentially accepted with its intention to assist users in identifying dangerous SMS messages. Hence, this would reduce the impact and risk. All three (3) respondents are agreed on the necessity of the proposed model as the following:

i. The proposed model is reliable in assisting users to detect malicious messages.
ii. The proposed model is feasible in daily use as life routine.
iii. It is a good thing to have this proposed model as mobile apps for daily use.
iv. The proposed model can increase awareness on spam's impact among users.
v. It is potentially to expand the usable of this proposed model to other platforms such as email and social media.

Details of this validation item is articulated as in Appendix F.

2.4 Mechanism Of Controlling Spam And Its Impact - Policy, Technical And Industry Approaches

Since the emergence of spam, in its initial form as email messages, there have been a lot of planned and implemented activities in order to curb the dissemination of this threat and its impacts. All levels of society or multi-stakeholders; individuals, government sectors and related private sectors such as telecommunication service providers and banking sectors have come together to execute dedicate procedure in terms of policy enforcement, technical and industry approach.

Unfortunately, this co-evolved threat has made almost all solution are not practical for certain time of period. Spam that always evolved and improved in its behaviour just to evade the anti-spam solution makes it hard to be totally diminished. The integration of mobile and Internet technology and the advancement of the traditional mobile phone to smart phone have made spam becoming even expanding in terms of its format (text to image, voice and video spam), platform (email to mobile and web page) and impacts (sale advertisement to criminal activity). At these days, its adverse effects do not only involve financial loss, but it is worsening to activity that includes breaking the law and regulation.

2.4.1 Organizations – Governments And Private Sectors

Spam has been flooded all over the world, which a country with a high ranking in spam lists typically has a high number of Internet connections (Natris, 2014). This is one of the notable findings from IGF. Spamhaus[5] discovered that the United States, China, and Russia are the top three (3) countries with highest spam rate in the first quarter of 2017.

Around the world, governments are taking measures to combat spam. For example, the Australia government established Australian Spam Act and Codes of Practice (2003) that covers email, instant messaging, SMS and MMS messages. Under this act, it is illegal for unsolicited commercial electronic messages to be sent to an address accessed in Australia. The legislation set up penalties of up to $1.1 million a day for repeat corporate offenders. The same law is enforced by the Canadian government with the establishment of Canadian

[5] Founded in 1998, Spamhaus Project is an international nonprofit organization that tracks spam and related cyber threats worldwide. http://www.spamhaus.org

Anti-Spam Act (2010). United State and European countries also implemented the same legislation via their own version of the law (n.a, 2012).

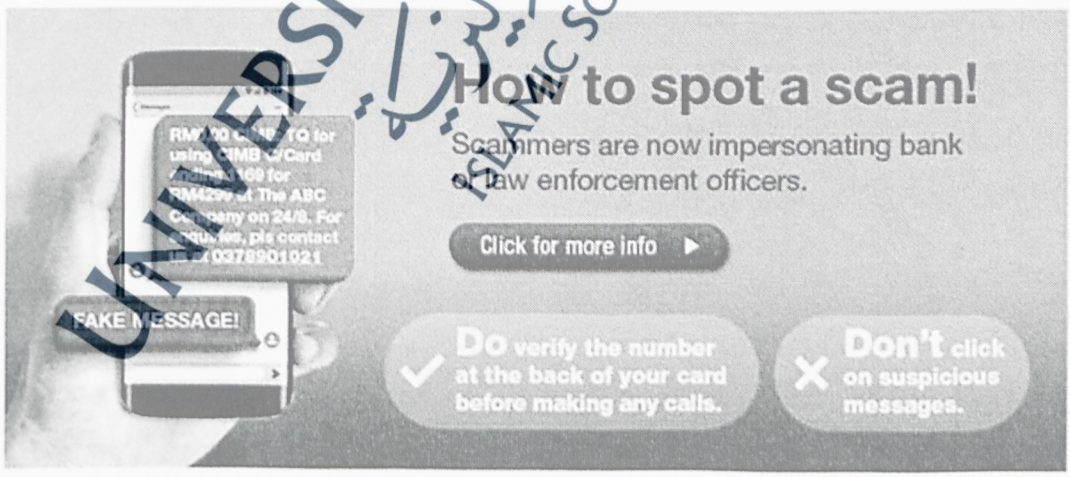
With the enforcement of these laws, it is easier for the government to combat spam and curb this threat. In China, 1,530 spammers have been caught for sending spam text messages (Farrell, 2014). Via this mass arrest, over 2,600 fake mobile base stations have been seized and 24 sites manufacturing illegal telecoms equipment has been shut down as part of a massive nationwide operation. The Chinese government also had identified 3,540 cases of suspected crimes, including one case where a Liaoning Province gang is suspected of sending out over 200 million spam messages.

Besides governments, this issue has been an agenda for non-government also to deal with spam. In the UK, mobile network operators have teamed up to stop and punish companies that are sending spam text messages (Rickhas, 2014). They also provide a service for mobile users reporting the spam problem which this center is operated on behalf of GSMA by Cloudmark, a company that specializes in messaging security software. The reported spam messages are aggregated and analyzed, providing operators with important details about the origin and size of the attack. Operators can then share the data in real-time to ensure they can all block and cut off spammers. Some service providers also take initiative in evaluating the best tool in combating the spam at their telecommunication gateway such as conducted by Warade, A.Tijare, & N.Sawalkar (2016) and Yan, Eidenbenz, & Galli (2009). A study by Singhal et al., (2013) claimed that non-content based filtering technique are largely employed by cellular network operators.

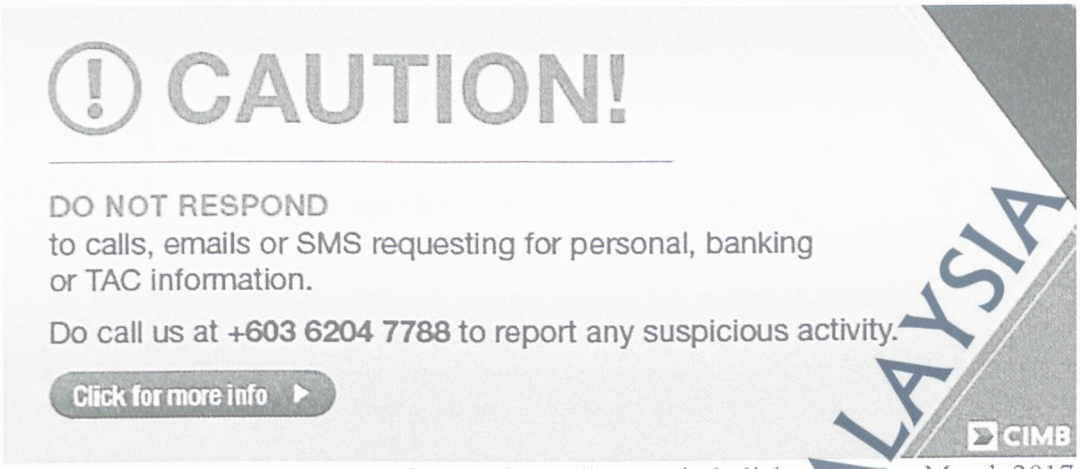
Another story of success is in Philippine where Globe Telecom, a telecommunication provider successfully blocked more than 165 million of spam messages in 2016, an average of 200,000 messages blocked per day. This is executed with the determination of an accelerated campaign versus unwanted spam or scam messages and aided by the use of a comprehensive plus fully automated mechanism to get rid its network of unwanted and unsolicited text messages (Telecom, 2017).

Besides legislative and regulatory measures, many governments produce educational material objectively to educate their citizens in spam

countermeasures. For example, the government of Canada produced a simple poster but informative and yet very comprehensible about spam avoidance (Canada, 2012). The Malaysian government is doing the same in educating the citizens about the harm of spam. Through collaboration with CyberSecurity, various type of education material has been distributed and via online, which include posters and videos to give awareness to the Malaysian, targeting multi-level of groups; kids, youth, adults, and organizations. In addition to that, to show the seriousness of Malaysian government to handle this spam issue, the Malaysian Communications and Multimedia Commission (MCMC)[6] has taken legal action against seven (7) content providers, compounded up to RM300, 000. These content providers are found flouted the law and acted irresponsibly in providing their mobile content services to consumers (n.a., 2015). Other than MCMC, there are other authority bodies that enforce and investigate complaints on content in the Internet, such as Royal Malaysia Police, Bank Negara Malaysia, Companies Commission of Malaysia and Ministry of Domestic Trade, Co-Operatives and Consumerism. Many authorized organizations in Malaysia also closely working together to disseminate the latest information such as technique that being used by scammers and potential targeted victims. For instance, as depicted in Figure 2.2, one of the Malaysian local banks is sharing the information on how scammers are impersonating bank or law enforcement officers to gain their trust. This type of awareness also has been considered as an informal training to users for being alert with fake services.



[6] https://skmm.gov.my/home



Source: https://www.cimbclicks.com.my, March 2017

Figure 2.2: Sample Of An Alert Notifications From Bank's Website

2.4.2 Individuals – Users, Consumers, And Researchers

Since that controlling spam is required combined efforts from all levels of community, besides organizations that include both governments and private sectors, individuals such as users, consumers, researchers or end-user organizations would be able to joint action in contributing to manage spam and its impacts.

There are many records of works that have been done in spam detection. Various mobile apps also have been developed that users can get the software to be installed on their phone. Even though the effectiveness of detection is giving a different result, at least there are efforts from users that realize and aware about this issue. Some comparison of top twelve (12) mobile apps effectiveness in SMS spam detection in Android is reported in Narayan & Saxena (2013). These apps include AVG, SMS Blocker, Quickheal, AntiSpam SMS, Numbercop, Private Box, SMS Filter, The Call, Postman, SMS Spam Blocker, smsBlocker and Spam Blocker. The testing also covered to test these mobile apps for email spam.

Prior to develop these apps, there are elements of feature extractions that researchers need to be considered to be applied in their proposed algorithm. For SMS spam environments, content and non-content features usually become the potential features and machine learning algorithms noticed as the most

outstanding classifier for this task. The summary review for choices of feature extractions and classification mechanisms that have been studied in SMS spam filtering is tabulated in Table 2.1 and simplified as depicted in Figure 2.3. This is not intended to be all-inclusive but rather to illustrate the most widely applied of feature extraction and classifiers in SMS spam filtering.

Table 2.1: Related Works In Spam Filtering For Previous Applied Feature Extractions And Classifiers

SMS Spam Filtering: Type of Feature Extractions		
Type	**Description**	**References**
Content	A content feature refers to message context that usually consists of spam keywords, URL links, monetary value, special characters, emotion symbols and function words.	Jain & Gupta (2018); Waheeb & Ghazali (2017); Suleiman & Al-Naymat (2017); Ezpeleta, Garitano, Zurutuza, & Hidalgo (2017); Choudhary & Jain (2017); Warade et al. (2016); Sulaiman & Jali (2015); Al-Hassan & El-Alfy (2015); Mujtaba & Yasin (2014); Karami & Zhou (2014b); Karami & Zhou (2014a); Warade, Tijare, & Sawalkar (2014); Song, Ye, Du, Huang, & Bie (2014); Mosquera et al. (2014); Ricknas (2014); Xia, Fu, & Zhou (2013); Almeida, Hidalgo, & Silva (2012); Tan, Goharian, & Sherr (2012); Mahmoud & Mahfouz (2012b); Muhammad Zubair Rafique & Abulaish (2012); Belén & Duarte-Figuriredo (2011); Mathew & Issac (2011); Khemapatapan (2010); Cormack, Hidalgo, & Sánz (2007); Hidalgo, Bringas, & Sánz (2003).
Non-content	Non-content features consider message metadata such as message length, the number of characters or words, white spaces, the number of terms, date, time and location wise. Other than that, blacklisted numbers and capital alphabet/word also considered as non-content features.	Suleiman & Al-Naymat (2017); Al-Hassan & El-Alfy (2015); Sulaiman & Jali (2015); Karami & Zhou (2014a); Karami & Zhou (2014b); Iyer & Shanthi (2013); Xu, Xiang, Yang, Du, & Zhong (2012); Yadav, Saha, Kumaraguru, & Kumra (2012); Alper Kursat Uysal, Gunal, Ergin, & Gunal (2012); Mathew & Issac (2011); Yadav, Kumaraguru, Goyal, Gupta, & Naik (2011); Yoon, Kim, & Huh (2010); M. Zubair Rafique & Farooq (2010); H. Zhang & Wang (2009); Cormack et al. (2007)

Table 2.1, continued

SMS Spam Filtering: Type of Classification Algorithms		
Type	**Description**	**References**
Machine Learning	Referring to learning algorithms which the classification executed with training phase prior to testing phase. There are evolutionary and non-evolutionary algorithms. Evolutionary algorithm uses mechanisms inspired by biological evolution such as AIS, Genetic Algorithm, and Ant Colony Optimization. While non-evolutionary algorithms is non-biological inspired such as SVM and NB.	Evolutionary algorithms: Al-Hassan & El-Alfy (2015); Sulaiman & Jali (2014); Mahmoud & Mahfouz (2012b) Non-evolutionary algorithms: Choudhary & Jain (2017); Zalpuri & Arora (2015); Yadav et al. (2012); Almeida et al. (2012); Tan et al. (2012); Xu et al. (2012); Belém & Duarte-Figueiredo (2011); Mathew & Issac (2011); Junaid & Farooq (2011); Yadav et al. (2011); Khemapatapan (2010); H. Zhang & Wang (2009); Cormack et al. (2007)
Linguistic	Linguistic analysis involves an analysis of language form, language meaning, and language in context. There are a few type of analysis that have been applied to distinguish spam messages such as Linguistic Inquiry and Word Count (LIWC), Latent Dirichlet Allocation (LDA) and Message Linguistic Analysis (MELA)	Onanuga (2017); Almeida, Silva, Santos, & Hidalgo (2016); Eshmawi (2015); Karami & Zhou (2014a); Karami & Zhou (2014b); Mosquera et al. (2014)
Challenge response	Some researchers applied challenge response to detect spam. One of the prominent methods is using CAPTCHA or Completely Automated Public Turing test to tell Computers and Humans Apart. This challenge aimed to distinguish between human users and computer programs automatically.	Yoon et al. (2010); Shirali-Shahreza & Shirali-Shahreza (2008)
Graph pattern	Include ontology pattern which formalize the information (context, data source), and generate mapping rules in accordance with the structure of ontology model of mobile phone spam	Muhammad Zubair Rafique & Abulaish (2012); Zhao, Zhang, Wang, & Liu (2012); Cao, Nie, & Liu (2011)

Although the feature extraction is divided into content and non-content, some research combined both techniques in their study of spam detection, such as in; Choudhary & Jain (2017); Suleiman & Al-Naymat (2017); Sulaiman & Jali (2015).

There are also other researchers who are studying and review the existing filtering mechanism for SMS spam which can be found in Abayomi-Alli et al. (2015); Abdulhamid et al. (2017); Lota & Hossain (2017); Foozy, Ahmad, & M.A. (2014). This review of spam detection is continuously expanding and various versions of review forms are introduced. For instance, authors in Abayomi-Alli et al. (2015) proposed three (3) types of spam extractions; content, non-content and listings (blacklist and whitelist) instead of only two (2) methods that are proposed in this study. Despite of these differences, there is no standard rule in structuring the taxonomy of this information.

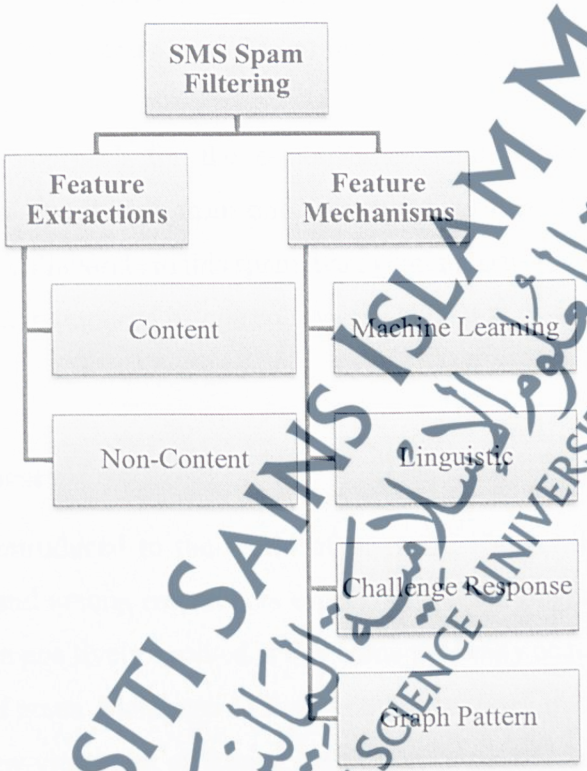


Figure 2.3: Simplification Of SMS Spam Filtering Methodology

Researchers in this domain always have been striving to update their work continually to find the accurate solution in helping mobile users control the spam problem. Some of their works even have been tested on smart phone using Android as a mobile platform and also done in multi-language besides English like Turkish, Chinese, Malay and African language to meet the needs of mobile users globally. This research can be found in Onanuga (2017); Balubaid

et al. (2015); Sethi & Bhootna (2014); Mohd Foozy et al. (2014); Yadav et al. (2012); Griesel & Fourie (2012) and Nuruzzaman, Lee, & Choi (2011).

In another aspect of spam detection, there is an industrial company that produced software for users understanding better about spam. For instance, Constant Contact developed a tool for email users that is used prior sending out email messages. This tool basically checking the content of email messages with the objective to minimize its possibility being blocked as spam at the recipient end (ConstantContact, 2010).

It is proven that every type of spam is unique and various mechanisms for anti-spam solutions are required for a different type of spam. Every mechanism demonstrated a different level of accuracy for its classification or detection rate.

The approach for the establishment and enforcement of anti-spam mechanisms also differs from one country to another. But it is important for countries to collaborate in this spam management issue to make it more effective and efficient in terms of its controlling mechanisms.

2.5 Danger Theory

As firstly introduced to the world of research, Danger Theory has created a controversial demand among researchers especially immunologists. Throughout time, this theory has been positively evolved and become gradually acceptable and applied in various domains of areas. Matzinger (1994) who has initiated this theory back in 1994 has suggested a new viewpoint of human immune systems which is the possibility the need to detect and protect against danger. This is contradicting to the classical belief that the immune system's primary motivation is the need to discriminate between self and non-self. Then, a further explanation of essential differences between the expanded Self Non-Self model and Danger model including signals that initiate immune responses is discussed in Matzinger (1998). In 2002, Matzinger further clarified and enhanced the basic version of Danger Model (Matzinger, 2002).

An auxiliary research as to fit in with the computational intelligence ambiance is done by a group of researchers from the United Kingdom. Begin with theoretical understanding, back in 2003, Aickelin et al., presented a Danger Project which this

study objectively to create the next generation intrusion detection system based on Danger Theory. They believe that this theory is the key that will unlock the true potential of AIS and the mechanism has the advantage of scanning intrusions at an early stage. This particular theory is further developed into, Dendritic Cell Algorithms (DCA) (Greensmith, 2007) and Toll-Like Receptors (TLR) algorithms (Twycross, 2007).

According to Greensmith et al. (2009) and Greensmith, Aickelin, & Tedesco (2010), DCA differs from other AIS algorithms for the following reasons:

i. multiple signals are combined to assess the current context of the environment;
ii. the correlation between context and antigen leads to the detection of anomalies;
iii. there is no pattern matching to perform the anomaly detection as conducted in negative selection;
iv. cells of the innate immunity are used as inspiration, not the adaptive immune cells and hence no dynamic learning is required; and
v. no dynamic learning is attempted, which means DCA does not rely on training data but instead domain or expert knowledge is required to predetermine the mapping between input and output signals from a particular instance to the categories of signals used in DCA and also the level of malicious (Gu, Greensmith, Oates, & Aickelin, 2010).

2.5.1 Abstraction View Of Danger Theory

As aforementioned in the previous section, Danger Theory is the most recent development in AIS. This theory that is found by Polly Matzinger in 1994, which focusing on what is dangerous instead of self and non-self discrimination idea. It implies that the immune system can discriminate between danger and non-danger (Matzinger, 1994). Danger Theory suggests that foreign invaders that are dangerous stimulate the production of cellular molecules (danger signals) by commencing cellular stress or cell death (Matzinger, 2002; 2007). The prominent Danger Theory that has been established is the Dendritic Cell Algorithm (DCA), which is signal processing algorithm and inspired by the behaviour of dendritic cells (Greensmith, 2007).

Dendritic cells or DCs are the fundamental possession of Danger Theory, which also is an innate immune system, and truly an intrusion or anomaly detection agent in the human body. They are Antigen Presenting Cell (APC) that responsible to digest antigen material and forward it on the cell surface to the T-cells of the immune system. This APC is acting as messengers between the innate and the adaptive immune systems.

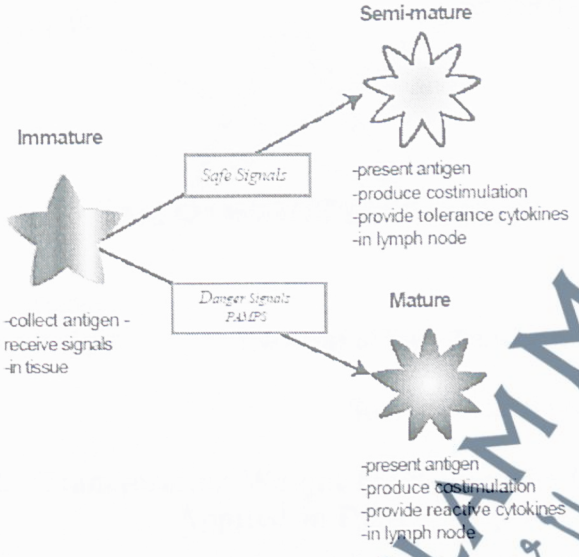
Wieder (2003) specified that DCs are capable of capturing antigens, processing them, and presenting them on the cell surface along with appropriate co-stimulation molecules. For this reason, DCs are unique APCs and have been referred as professional APC. In addition to that, only DCs have the ability to induce a primary immune response in resting naïve T lymphocytes. This makes DCs critical in the establishment of immunological memory too.

These DCs are possessing in the body tissues and gather antigen and other (danger) signals that give a picture of the current condition of the tissues. This representation of the current state figures out if the antigen has been gathered in a safe or dangerous context, and causes DCs to change into a semi-mature or mature state. The mission of the DCs are to distinguish antigens as being either benign (harmless) or malignant (harmful) in nature (Greensmith, Aickelin, & Cayzer (2010); Greensmith, Aickelin, & Tedesco (2010)).

Aickelin & Greensmith (2007) emphasized that prior the transformation of immature DCs to either semi-mature or mature state, there are three (3) types of molecules in term of signal released exclusively by pathogens to indicate the dangerous or safe of its context. Pathogens Associated Molecular Patterns (PAMPs) signal signify abnormal behaviour which highly indicates of an anomaly. Then, danger signal is less than PAMPs and safe signal released to indicate safe context. While inflammation is categorized as molecules of an inflammatory response to tissue injury and it also acts as a natural amplifier for all other signals.

The signals that migrated to the lymph node are divided into two (2) types of signals as regards to the degree or concentration of danger detected. Apoptotic alerts or semi-mature brings the safe signal, while necrotic alerts bring the mature signal. Semi-mature indicates a 'safe' context and mature

indicates a 'dangerous' context. These signals are a reflection of the state of the surrounding. A general sign of system distress called as inflammatory signals (Greensmith, Aickelin, & Cayzer, 2010); (Greensmith, Aickelin, & Tedesco, 2010). All of these signals can be depicted as in Figure 2.4 below.



Source: Greensmith, Aickelin, & Tedesco (2010)

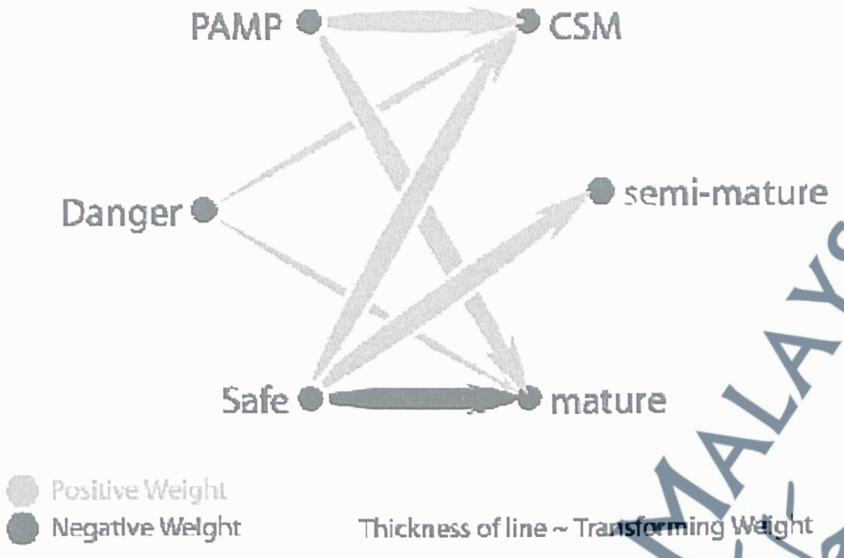
Figure 2.4: The Transformation Of Immature DCs

Greensmith, Aickelin, & Cayzer (2010) elaborated that the generated output signals are at a certain concentration and proportional to the received input signals. The calculation can be measured as in equation (2.1).

$$Oj = \sum_{i=0}^{2} (Wij \times Si) \qquad (2.1)$$

where O*j* are the output signals, S*i* is the input signals and W*ij* is the transforming weight from S*i* to O*j*.

The transforming weight is referring to Figure 2.5 and it showed a schematic diagram of the signal processing equation used by every DC to fuse input signals and derive output signals. The thickness of the line depicts the transforming weight, the thicker the line, the more weight is consumed in the assessment. Inflammation signals are not shown in the figure.



Source: Aickelin & Greensmith (2007)

Figure 2.5: The Transforming Weight From Input To Output Signals Applied In DCA

This Figure 2.5 can be described in table form as tabulated in the following Table 2.2.

Table 2.2: The Derivation And Interrelationship Of Weights In The Signal Processing

Signal	PAMPs	Danger	Safe
CSM	W1	W1*0.5	W1*1.5
Semi-mature	0	0	1
Mature	W2	W2*0.5	-W2*1.5

Source: Greensmith, Aickelin, & Cayzer (2010)

With reference to Figure 2.5 and Table 2.2, the values of the PAMPs weights are used to create the all other weights relative to the PAMPs weight. For instance, W1 is the weight to transform PAMPs signal to the co-stimulation or CSM output signals and W2 is the weight to transform the PAMPs signal to the mature output signal. The negative value weight for Safe signal to transform to the mature state depicted that it is negative to be a mature state or harmless outcome. This sum is repeated three (3) times, once per output signals; CSM, semi-mature and mature. This is to calculate the interim output signal values for these three (3) signal values and cumulatively summed over time. Incrementing

CSM output signal is an important feature of the DCA algorithm, as it provides a limit to the time spent by the DCs sampling the data. The CSM signal indicates that the DCs has collected sufficient signals to make a decision for the maturation status, either semi-mature or vice versa.

The actual values used for the weights can be user defined, though the relative values (as in Table 2.2) determined empirically are kept constant. These signals are used to assess the state of DCs upon termination of the sampling phase of a DC life span.

The level of maturity is determined by the data of antigen collected by DCs. This measurement is calculated using the Mature Context Antigen Value (MCAV) which this is the mean value of context per antigen type, or in another word, determine the intensity or degree of the detected danger (Greensmith, 2007). This can be measured by the following equation:

$$MCAV\ (antigen\_type) = \frac{mature\_count}{antigen\_count} \qquad (2.2)$$

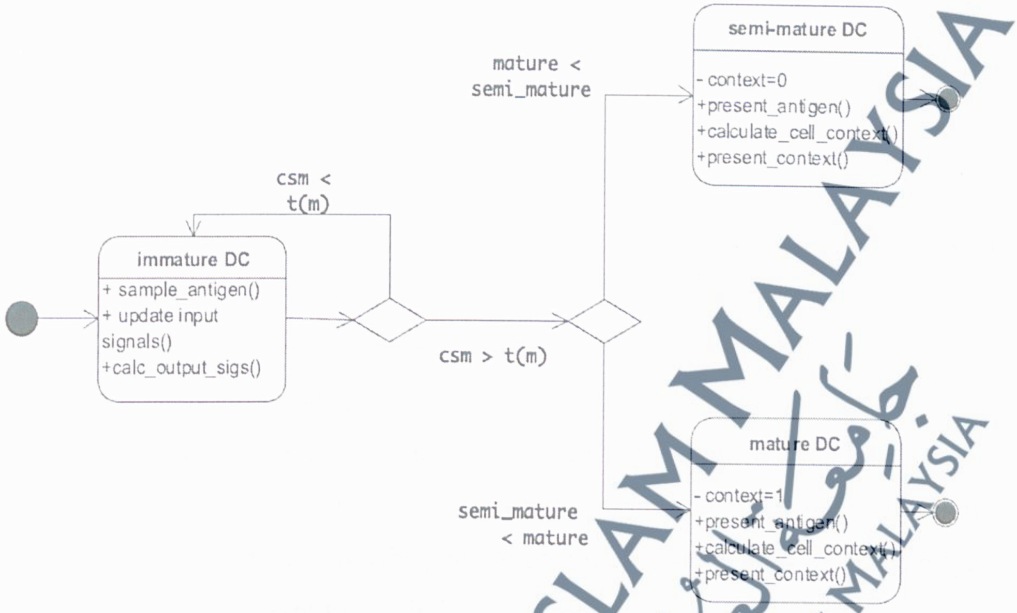
As clarified in Greensmith et al. (2009), if this theory of algorithm performs as intended, the closer this value to 1, the greater the probability that the antigen is anomalous. This MCAV value is used to assess the degree of the anomaly of a given antigen.

2.5.2 DCA As An Information Data Fusion, Signal Processing, And Correlation Algorithm

All the basic processes that DCA involved in any anomaly detection process can be simplified as in Figure 2.6 below followed by an elaboration of how immature DCs transform into either mature or semi-mature state. This figure shows the different possible states over a DC life span, where t(m) represents the migration threshold of the cell and CSM is the co-stimulation output signal. Biologically, the processes on the left occur in the tissue and on the right occur in the lymph node.

This explanation is gathered from collective references but not limited to these, which have become the foundation theory of this research. These can

be found in Greensmith, Aickelin, & Tedesco (2010); Greensmith et al. (2009); Aickelin & Greensmith (2007); Greensmith (2007); Aickelin & Greensmith (2007) and Aickelin et al. (2003).



Source: Aickelin & Greensmith (2007)

Figure 2.6: DC State Transition Overview Diagram

i. While in the immature state, the DC has three (3) functions that occur asynchronously.

- Antigen sampling – the DC collects antigen from an external source and places the antigen in its own antigen storage data structure;
- Update input signals – the DC collects values of all input signals present in the signal storage area;

Determine interim output signals – at each iteration, every DC calculates three (3) temporary output signal values (refer Equation (2.1)) from the received input signals, with the output values then added to form the cell's cumulative output signals.

ii. Upon initialization, each DC is assigned what is termed a migration threshold, t(m). The value of anomaly threshold separates normal and

abnormal antigen. Output signal that measured greater than this value indicated as malicious context.

iii. Prior to a recalculation of the output signals, the achieved value of the CSM output signal as compared to the cell's migration threshold value, t(m). If the value of the CSM signal is greater than the migration threshold, then the DC ceases sampling signals and antigen and is then transferred to a separate compartment and is assigned a new state (semi-mature or mature).

iv. Immediate replacement for the removed cell from the population with a new one is required to fix the population size at a static level. If the cell does not surpass its pre-defined migration threshold, t(m) it continues sampling and the output signals accumulate. In implementations of the DCA, each DC is assigned a random number (within a specified range) for the migration threshold, t(m). This ensures that across the population, the DCs sample signals and antigen over different time windows.

v. The remaining two (2) output signals of the DC are assessed once this threshold is exceeded. At this point, the value of the semi-mature output signal is compared to the value of the mature output signal. The context of the cell is assigned as that of the greatest output signal value.

vi. The presented antigen are accompanied by the context of the cell (0 for semi-mature and 1 for mature) and are recorded. This information is used following the processing of all input data to calculate the MCAV anomaly coefficient value for each type of antigen (refer Equation (2.2)). The MCAV calculation returns a coefficient value between 0 and 1. Values closer to 1 indicate that the antigen type has a higher probability of being anomalous. This completes the pairing of 'suspect' antigen with 'evidence' from signals, based on the consensus opinion of the DC population over time.

```
input : signals from all categories and antigen
output: antigen plus context values
initialiseDC;
while CSM output signal < migration Threshold do
    get antigen;
    store antigen;
    get signals;
    calculate interim output signals;
    update cumulative output signals;
end
cell location update to lymph node;
if semi-mature output > mature output then
    cell context is assigned as 0 ;
else
    cell context is assigned as 1;
end
kill cell;
replace cell in population;
```

Source: Greensmith (2007)

Algorithm 2.1: Generic DCA Algorithm That Depicts The Entire Process Of Danger Measurement

2.5.3 Characteristic Of Danger Theory Applied In Computational Intelligence

This research is employed a Dendritic Cell Algorithm (DCA) from Danger Theory, that has the ability to combine multiple signals to assess the current context of the environment and also to asynchronously sample another data stream (antigen). The correlation between context and antigen is used as the basis of anomaly detection in this algorithm (F Gu, Greensmith & Aickelin, 2008). Anomaly detection or danger classification is performed through the correlation of antigen with signals (Greensmith, Aickelin & Twycross, 2009).

In addition to that, other researchers Chelly & Elouedi (2015); Pradeu & Cooper (2012) and Danziger & De Lima Neto (2011) claimed that from their simulation findings, advantages of Danger Theory are identified as having low count of false positive, high number of true positive, no training data required though only minimal amount of data pre-processing is needed, ability to solve a

non-real time problem, empirically lightweight in terms of running time due to simple linear function of equation.

To have a better understanding of the DCA application as described in Section 2.5.1 and 2.5.2, the various paper of other researchers have been referred to be studied and examined in detail. DCA application has been employed in various fields of computational intelligence and part of the research evidence is tabulated in Table 2.3.

From this Table 2.3, it is clearly shown that DCA has been mostly applied in intrusion detection field in the environment of wireless network, mobile ad-hoc network and cloud computing networking systems. It also demonstrated that Danger Theory does not acquire large training set that make it less complicated (Ge, Liang, Chen, & Zhang, 2015). It is also shown that Danger Theory is well-applied in detection of malware. Foremost, this theory are capable to measure the risk probability of detected anomaly quantitatively as demonstrated for spam classification (Al-Hassan & El-Alfy, 2015) and assessment of risk for cyber threat (Mihai-Gabriel & Patriciu, 2014). Danger Theory also proven as well-consolidated to be integrated with other prominent algorithm such as Naïve Bayesian (NB) and Support Vector Machine (SVM).

Table 2.3: Application Of DT In Computational Intelligence

Domain of application	Description	References
Intrusion detection	In the current paper, a bio-inspired method is introduced, namely the Cooperative-based Fuzzy Artificial Immune System (Co-FAIS) for intrusion detection in a wireless network. It is a modular-based defense strategy derived from the DT of the human immune system. The agents synchronize and work with one another to calculate the abnormality of sensor behavior in terms of context antigen value (CAV) or attackers and update the fuzzy activation threshold for security response. The model is subsequently compared against other existing soft computing methods, such as Fuzzy Logic Controller (FLC), AIS, and Fuzzy Q-Learning (FQL). The proposed method improves detection accuracy and successful defense rate performance against attacks compared to conventional empirical methods.	Shamshirband, Anuar, Kiah, & Rohani (2014)
Intrusion detection	Authors proposed two (2) applied AIS for intrusion detection using the *KDD Cup'99* database. The first one is based on the DT using DCA and the second is based on negative selection algorithm (NSA). Both algorithms are implemented in Java in NetBeans IDE. Receiver Operating Characteristic analysis or ROC is performed to evaluate the performance. The results for the DCA are quite encouraging but in contrast, the NSA did not provide conclusive results. It emits a large number of false alarms in contrast to the DCA algorithm whose false alarm rate is around 0. Authors also noted that NSA has difficulty in managing a large data set, which is a serious drawback, given the current size of database computer systems.	Zekri, Souici, & Meslati (2014)
Intrusion detection	A mobile ad hoc network (MANET) is an open wireless network of mobile, decentralized, and self-organized nodes with limited energy and bandwidth resources. The MANET environment is vulnerable to dangerous attacks, such as flooding-based attacks, which paralyze the functionality of the whole network. This paper introduced a hybrid intelligent algorithm, for protecting the MANET with effective security and network performance. This objective is fulfilled by inspiring and integration of anomaly detection of DCs in DT and fuzzy logic theory to introduce a Dendritic Cell Fuzzy Algorithm (DCFA). DCFA combines the relevant features of DT-based AIS and fuzzy logic theory-based systems. DCFA is verified using QualNet v5.0.2 to detect resource consumption attack. The results show the efficient capability of DCFA to perform the detection operation with high network and security performance.	Abdelhaq, Alsaqour, Ismail, & Abdelhaq (2015)
Intrusion detection	The paper proposed a new model by applying bio-inspired theory into intrusion detection system in cloud computing networking systems. The core detection module use DCA which does not require large training sets and the knowledge of normality and anomaly is acquired through a machine learning approach, which is the second improvement to achieve a more accurate detection rate and a low false negative rate, that enhance the whole performance of the entire detection system	Ge, Liang, Chen, & Zhang (2015)
Intrusion detection	This paper analysed the built up of an immunity model in the network against the attacks. The accuracy, time analysis immune memory strength and fittest all level, parameters are studied using two (2) algorithms; immune networks algorithm and DCA.	Kumar (2015)
Intrusion detection	In this work, authors proposed a novel intrusion detection system that is based on the trust value of node and classifier inspired by DCA theoretically. Decision phase generated a list of permanent and temporary blocking nodes according to the trust level. The expected outcome of the system will be the effective detection of any intrusion with reduced false alarm rate, misclassification rate and improved detection accuracy.	Singh & Bedi (2015)

Table 2.3, continued

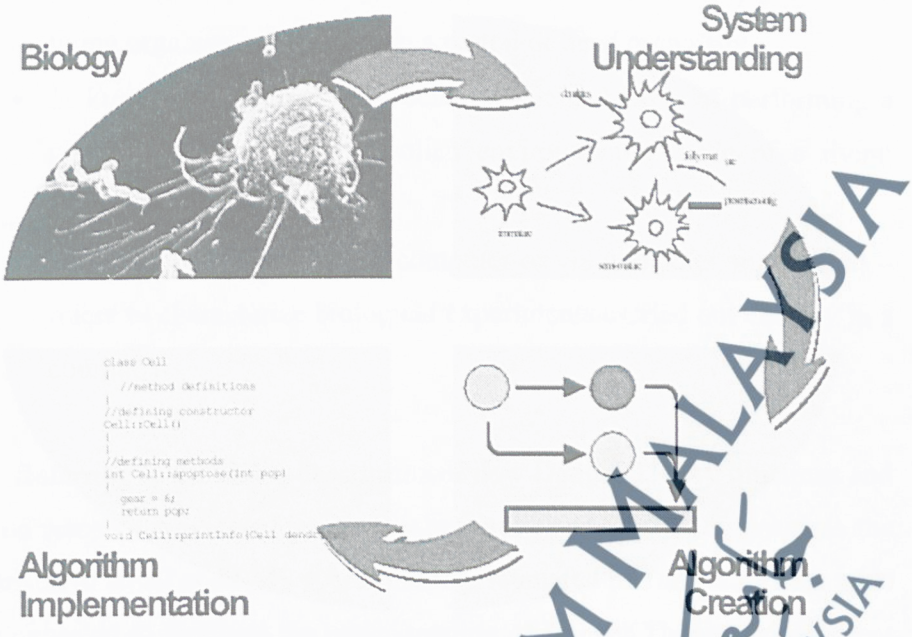
Domain of application	Description	References
Intrusion detection	This paper particularly applied DT to implement Danger Theory Based Model for Network Security Risk Assessment or DTMNSRA that evaluate the network security risk quantitatively in real-time. The higher attack frequency and intensity, the more dangerous the network faces. In DTMNSRA, local hosts can detect the type of disease that they are suffering from and identified the severity of the disease. This induced the generation of danger signals by simulating cellular distress or cell unnatural death in DTMNSRA The architecture of DTMNSRA is mainly composed of sensors and a risk assessment center. The sensor is located on each host, and it is in charge of the detection of network attacks. The functions of the risk assessment center include two (2) aspects: bacterin distribution and network security risk assessment.	Sun (2011)
Intrusion detection	A mobile ad hoc network (MANET) is a set of mobile, decentralized, and self-organizing nodes that are used in special cases, such as in the military. MANET properties render the environment of this network vulnerable to different types of attacks, including black hole, wormhole and flooding-based attacks. Flooding-based attacks are one of the most dangerous attacks that aim to consume all network resources and thus paralyze the functionality of the whole network. The objective of this paper is to investigate the capability of a DT based artificial immune algorithm called the Mobile Dendritic Cell Algorithm (MDCA) to detect flooding based attacks in mobile ad hoc network (MANET). The MDCA applies the DCA to secure the MANET with additional improvements. The MDCA is tested and validated using Qualnet v7.1 simulation tool. This work also introduces a new simulation module for a flooding attack called the Resource Consumption Attack (RCA) using Qualnet v7.1. The results highlight the high efficiency of the MDCA in detecting RCAs in MANETs.	Abdelhaq, Alsaqour, & Abdelhaq (2015)
Anomaly detection	One of the problems in the computer security system is port scanning attack. There are several detection systems have been developed to find out the occurrence of port scanning attack, one of them is anomaly detection method. In this paper, authors design a simple implementation of anomaly detection system based on DCA. To determine a reviewed process tends to be anomalous, anomaly threshold coefficient is defined. The calculated value of anomaly threshold, 0.4759933 is quite valid and representative in order to determine the nature of anomaly of a process. Based on the test result, Nmap process which has 0.6164136 as the average MCAV value can be classified as an anomaly process within the host computer. Meanwhile, three (3) other reviewed processed; Bash, SSH, and SCP always have the average MCAV values below the defined anomaly threshold value, so these can be classified as normal processes.	Anandita, Rosmansyah, Dabarsyah, & Choi (2015)
Classification (spam)	As mobile devices continue to evolve, the volume of hacking activities targeting them also increases drastically. Receiving SMS spam is one of the common vectors for security breaches. This paper explores Danger Theory and applied it in SMS spam filtering. The filtering model using a combination of most relevant features and fusing decisions using two (2) machine learning algorithms, NB and SVM together with DCA. The required signals needed in DCA are also generated by NB and SVM. The performance has been evaluated on two (2) different sets of SMS spam data and via simulation showed that significant improvements can be achieved in the overall accuracy, recall, and precision of spam and legitimate SMS messages with the proposed DCA-based model. From the experiments, different parameters value are required for different set of data in achieving the optimum results. Hence, several experiments need to be conducted to identify the best value for number of DCs, antigen multiplier and signal weights.	Al-Hassan & EL-Alfy (2015)

Table 2.3, continued

Domain of application	Description	References
Malware detection	This paper proposed a new framework for malicious code immune mechanisms, Network Malicious Code Immune Model. The DCA successfully simulated aspects of immune mechanisms possessed by biological systems to achieve an immune effect against malicious code in computer networks. This paper demonstrated in detail the feasibility of the imbalanced Support Vector Machines (SVM) method in optimizing the immunization program output data. It also verifies experimentally that imbalanced SVM can optimize the output of the malicious code immune system by removing glitches from the outputs, making it easier to identify the threshold of malicious code immunization.	L. I. Peng & Ruchuan (2015)
Malware detection	This paper applied DT in suggesting a 2-layer network model for simulating virus propagation through Bluetooth and SMS. This method effectively restrain virus propagation in a large-scale network with its four (4) phases model; danger capture, antigen presentation, antibody generation and antibody distribution.	M, Uma & Kumar (2014)
Risk assessment for cyber threat	Authors of this paper suggested DT as one of the important property in Early Warning Systems (EWS) based on intelligent threat assessment. The proposed approach is using an intelligent method of risk assessment for calculating the probability of certain cyber-attacks that are about to happen. This will be calculated periodically and when certain limits are exceeded, the proposed system will trigger an early warning so that human operators are prepared. As proof of concept, the proposed system is executed a risk assessment compared with Neural Networks. Results showed that the proposed system is competent to calculate the risks explicitly.	Mihai-Gabriel & Patriciu (2014)

2.5.4 The Conceptual Framework Or *In Silico* Processes For Danger Theory Application In The Field Of Computational Intelligence

In 2005, Stepney has suggested a Conceptual Framework prior to successful designing and developing AIS algorithm especially for the second generation of AIS. Authors proposed that bio-inspired algorithms are best developed and analysed in the context of a multidisciplinary conceptual framework that provides for sophisticated biological models and well-founded analytical principles. The developed theory usually is tested with a real world case study to verify the computational algorithms (Stepney et al., 2005). This framework takes into account the interrelationship between immunology and computer systems.



Source: Greensmith, Aickelin, & Cayzer (2010)

Figure 2.7: A Flow Diagram Of The Abstraction Process Used In Designation Of Theoretical Immunology And Computational Algorithms

Based on Figure 2.7, this methodology employs an iterative approach to the creation and testing of novel immune-inspired algorithms and consists of four (4) stages that are identified as:

- observation of the biological experimentation;
- constructed computational models;
- developed, implemented and studied the biological abstraction as algorithms; and
- applied the algorithm to a specific problem, with feedback for refinement.

This framework has been further transformed into structured process introduced by Figueredo, Siebers, Aickelin, & Foan (2012) as depicted in Figure 3.2. This approach can be applied in problem-specific investigation that utilized immunology field as the research basis.

Besides Conceptual Framework, there is terminology of broad study categories for biological experiments. They are known as *in vivo*, *in vitro* and *in silico;* in Latin and differentiate as the following (n.a., 2012b):

- *In Vivo* (within the living) – refers to examination using a whole, living organism as opposed to a partial or dead organism;
- *In Vitro* (within the glass) – refers to the technique of performing a given procedure in a controlled environment outside of a living organism; and
- *In Silico* (performed on the computer or via computer simulation) – refers to characterize biological experiments carried out entirely in a computer.

Referring to the above description of how Danger Theory functions and based on research conducted previously in spam classification, herewith is the comparison in email and SMS spam filtering, tabulated in Table 2.4. This table also is expected to facilitate the understanding of Danger Theory employment for this particular research, assessing the risk level of spam messages.

From Table 2.4, biological properties from Danger Theory are identified and mapped with the spam's environment, both for email and SMS messages. All the potential elements in the theory are illustrated and clarified in spam's environment which is delineated and later is applied in the model design (Chapter 3). For instance, antigen is a toxin or substance which induced an immune response in the body. In spam's environment, this is portrayed by spam messages that may cause potential risk or impact to mobile phone or computers. Other significant biological properties are illustrated in detail in the following Table 2.4, aligned with its characteristics in spam's environment.

Table 2.4: The Properties Of Danger Theory Applied In Text Messages Spam Filtering For Email and SMS Format

Properties in Biology (Danger Theory)	Abstraction used in Dendritic Cell Algorithm (DCA)	Applied Designation	
		Email Spam	SMS Spam
Significance and Importance	Due to limitations and drawbacks in classical AIS algorithm	Authors proposed a predictive model to classify email spam based on DCA. They used 3 different machine learning algorithm to produce the input signals, which are kNN (PAMPs signal), NB (danger signal) and SVM (safe signal). As for the features that to be extracted, they analysed the email header and message body. Term Frequency-Inverse Document Frequency (TF-IDF) is utilized in assigning a weight for each term.	Authors applied the same concept of email spam classification for detection of spam in SMS messages. Again, Naïve Bayesian (NB) and Support Vector Machine (SVM) are used to produce input signals, whereby PAMPs signal are generated when both classifiers agree that the SMS message is a spam. Feature extracted include URL link, spam words, emotion symbols, special characters, message metadata and function words or grammatical words. The combination of spam words and metadata giving the highest accuracy for all three (3) classifiers.
Antigen	Substances that can initiate adaptive immune response	Email messages	SMS messages
Dendritic cells or DCs	An antigen presenting cell that collects and process information of surrounding that need to be presented to T-cells for further action (suppress or reactivate)	Filtering mechanism (feature extraction using keywords, TF-IDF)	The model itself
T-cells	The main agent that will suppress or activate the immune response	Not Applicable	Not Applicable
DC stop sampling and migrated to lymph node	Reached lifespan. Value of migration threshold where the DC migrated from tissue to lymph node	Assigned as 0.39	No value of anomaly threshold to be found
Maturity affinity – the binding degree between T-cells and antigen	MCAV - The measurement of danger concentration or possible impact intensity	MCAV > anomaly threshold → spam	MCAV > anomaly threshold → spam

Table 2.4, continued

Properties in Biology (Danger Theory)	Abstraction used in Dendritic Cell Algorithm (DCA)	Applied Designation	
		Email Spam	Email Spam
Immature DCs	Collect antigenic material and exposed to input signals	Raw email messages (prior pre-processing phase)	SMS messages
Mature DCs	Have an activating effect	True positive email as spam	True positive SMS as spam
Semi-mature DCs	Have a suppressive effect	True positive email as legitimate	True positive SMS as legitimate
PAMPs signal	A signature of abnormal behaviour; A high degree of confidence of abnormality associated with an increase in signal strength	Generated using the confidence level of a kNN algorithm	If SVM==spam & SVM=NB, then PAMPs is maximum of SVM & NB
Danger signal (necrotic)	Measure of an attribute which significantly increases in response to abnormal behaviour; A moderate degree of confidence of abnormality with increased level of signal, though at a low level strength can represent normal behaviour	Employed NB to generate danger signal	If SVM ≠ NB, Danger is average of SVM & NB
Safe signal (apoptosis)	A confident indicator of normal behaviour in a predictable manner or a measure of steady behaviour; Measure of an attribute which increases signal concentration due to the lack of change in strength	Utilized SVM to produce safe signal	Not applicable
Inflammatory cytokines	A signal which cannot cause maturation of a DC without the other signals present; A general signal of system distress	Receiving a large number of unsolicited email messages	Not applicable
References	Greensmith et al., (2009); Kim et al., (2010)	M.El-Alfy & Al-Hasan (2014)	Al-Hassan & EL-Alfy (2015)

2.6 Danger Theory Variants

The Danger Model was initially proposed by Polly Matzinger, an immunologist, who suggested a novel explanation of how the immune system works. She proposed Danger Model (Matzinger, 1994), which suggests that the immune system is more concerned with the damage than cell foreignness. The immune system is called into action by alarm signals from injured tissues, rather than by the recognition of non-self. In this model, the important feature is that danger or alarm signals should not be sent by healthy cells (Matzinger, 2002).

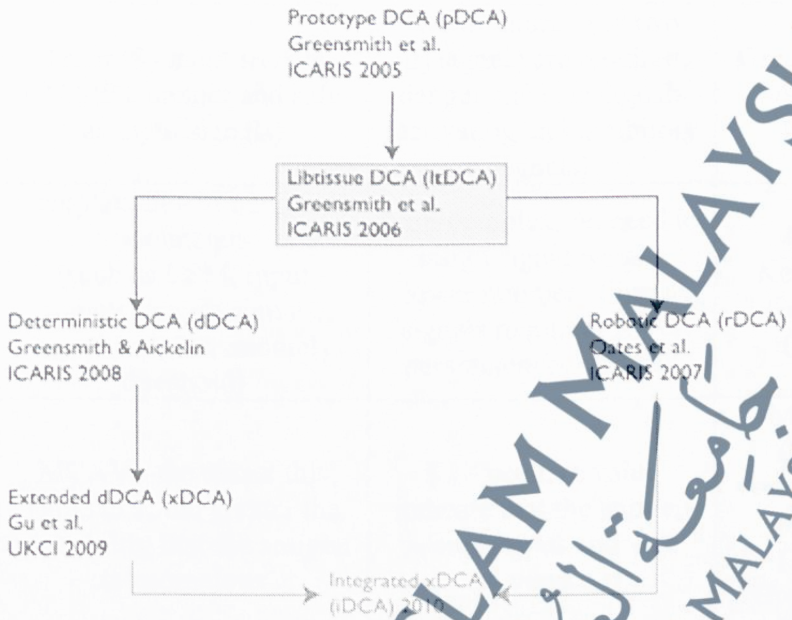
This idea then adapted in computational intelligence as Danger Theory that becomes known as the second generation of AIS. Later, in 2005, another two (2) hypotheses emerged from this theory, Dendritic Cell Algorithm (DCA) (Greensmith, 2007) and Toll-Like Receptor (TLR) (J Twycross, 2007). Both of these hypotheses are suggesting the linkage point that is integrated between innate and adaptive immune system.

If DCA focusing on DCs behaviour, TLR is inspired with the abstraction of two (2) population interacting cells, T-cells and naive DCs (Aickelin & Greensmith, 2007). Differences between DCA and TLR are discussed in Table 2.5.

Table 2.5: Theoretical Differences Between DCA And TLR

2nd Generation Of Artificial Immune Systems (AIS)	
Danger Theory (Polly Matzinger in 1994) In 1994, Matzinger proposed a theory in immunology called "Danger Theory" which states that the immune system is activated upon receipt of molecular signals which indicate damage or stress to the host rather than by pattern matching of 'non-self' versus 'self'.	
DCA (Dendritic Cell Algorithm) proposed by Julie Greensmith	**TLR (Toll Like Receptor algorithm) proposed by Jamie Paul Twycross**
Inspired by DCs characteristics and behaviours	Based on innate immune principals and tries to use and abstraction of T-cells, DCs, negative selection, tissue compartment and lymph nodes
Antigen and signal are considered separately	DCs collect antigens and at the same time process signal. This algorithm does not utilize different groups of input signals
DCA does not have an adaptive component and thus requires no formal training phase	TLR is completed in training and detection phases but it only uses normal samples in its training phase
Relies on the signal processing aspect by using multiple inputs and output signals	Emphasizes the interaction between DCs and T cells, only uses danger signal
References: Greensmith (2007); Greensmith, Whitbrook, et al., (2010); Aickelin & Greensmith (2007)	References: Jamie Twycross & Aickelin (2010); J Twycross (2007)

For DCA, the algorithm is keep evolved with numerous types of applications in the various field. A brief history and information on the DCA development can be found in Feng Gu, Greensmith, & Aicklein (2011).



Source: Feng Gu, Greensmith & Aickelin (2011)

Figure 2.8: Development Pathway Of The DCA

With detail examination of differences between DCA and deterministic DCA (dDCA), herewith is the theoretical differences (except for accuracy) between DCA and dDCA, tabulated in Table 2.6. These theoretical differences are applied for simulation analysis and further elaborated in Chapter 3.

Table 2.6: The Theoretical Differences Between DCA And dDCA

Feature / Attribute	DCA	dDCA	References
Naming	Dendritic Cell Algorithm, also known as classical DCA	Deterministic DCA	Greensmith & Aickelin (2008)
Input signals	Three (3) input signals (PAMPs, danger and safe as input signals)	At minimum only two (2) signals are required; danger and safe signals (activating and inhibitory signals)	Greensmith & Aickelin (2008)
Algorithm complexity	Complex; due to numerous parameters (such as CSM, input signals, output signals, signal weights, anomaly threshold)	Non-complex; no need to assign signal weight, lower number of input signals required, fewer parameters considered	Oates, Kendall, & Garibaldi (2008)
Anomaly metrics	MCAV - the closer this value to 1, the greater the probability that the antigen is anomalous	K_α – positive value indicate that the antigen is anomalous and vice versa	Musselle (2010); Greensmith & Aickelin (2008); Greensmith (2007)
Anomaly measurement (outcome of signal processing or data fusion)	*O[mDC] > O[smDC]* and MCAV > t_m - this is to indicate that spam message is malignant and MCAV result should return a value as the malignant message; or *O[mDC] < O[smDC]* and MCAV < t_m - this is to indicate that spam message is benign and MCAV result should return a value as the benign message.	$K_\alpha > 0$ and $K_\alpha > T_k$ - spam message is tagged as the malignant message; or $K_\alpha < 0$ and $K_\alpha < T_k$ - spam message is tagged as the benign message.	Greensmith & Aickelin (2008); Greensmith (2007)
Accuracy (technical perspective)	Acceptable rate (more than 80% of classification accuracy rate)	Higher than DCA and the risk concentration value (numerical) is finer-grained.	

A further discussion on how these variants (DCA and dDCA) function in assessing the risk intensity is articulated and demonstrated in Chapter 3.

```
input  : Antigen and Signals
output: Antigen Types and cumulative k values
set number of cells;
initialise DCs();
while data do
    switch input do
        case antigen
            antigenCounter++;
            cell index = antigen counter modulus number of cells ;
            DC of cell index assigned antigen;
            update DC's antigen profile;
        end
        case signals
            calculate csm and k;
            for all DCs do
                DC.lifespan -= csm;
                DC.k += k;
                if DC.lifespan <= 0 then
                    log DC.k, number of antigen and cell iterations ;
                    reset DC();
                end
            end
        end
    end
end
for each antigen Type do
    calculate anomaly metrics;
end
```

Source: Greensmith & Aickelin (2008)

Algorithm 2.2: Generic dDCA Algorithm That Depicts The Entire Process Of Danger Measurement

2.7 Spam Treated As A Threat With Risk

Conventional spam considered as a traditional threat with limited risk such as email that only can be accessed via computers in 1990's. Integration of the Internet and mobile technology then produced mobile spam that is more sophisticated threat with risk impact that is wider and more severe. SMS spam has been observed as one of the threats to mobile devices based on its malevolent impacts. This has been debated in Theoharidou et al., (2016) and Yeboah-Boateng & Amanor (2014).

This issue of threat and risk is seriously needed to be managed. Many policy and standard have been developed and establish to curb the issue of various threats mainly related to information security. The policy is required to guide a security administrator or risk evaluator in managing the risk. Some of the well known standards that is

successfully established is National Institute of Standards and Technology (NIST)[8], International Organization for Standardization (ISO)[9], PCI Security Standards Council[10], and Federal Emergency Management Agency (FEMA)[11].

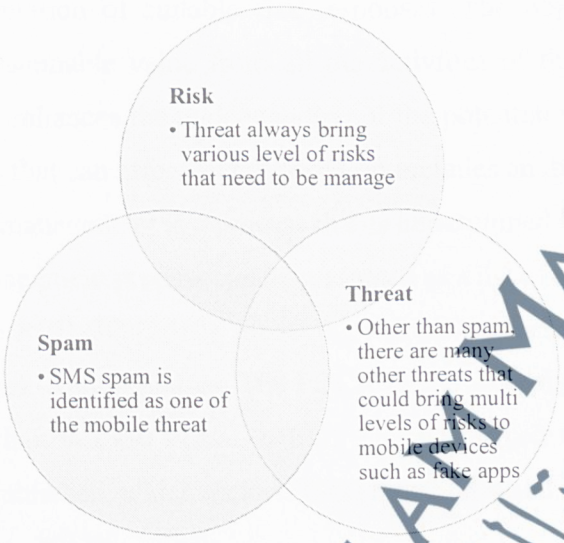


Figure 2.9: The Relationship Of SMS Spam As A Threat That Potentially Could Bring Risk To Mobile Devices And Other Installed Applications

2.7.1 Risk Management

By many standards, it is agreeable that risk is something that is an effect of uncertainty on objectives (Luko, 2013) whereas this

i. effect is a deviation from the expected, either positive or negative; and
ii. objectives can have different aspects (such as financial, information, and safety goals) and can apply at different levels (such as individual or organization).

The risk is often characterized by reference to potential events and consequences or a combination of these, associated with likelihood or probability of occurrence of the event.

Managing risk is an essential activity for enterprises of all sizes, including the individual. Enterprises that manage risks effectively will thrive

[8] https://www.nist.gov/
[9] https://www.iso.org/home.html
[10] https://www.pcisecuritystandards.org/pci_security/
[11] https://www.fema.gov/

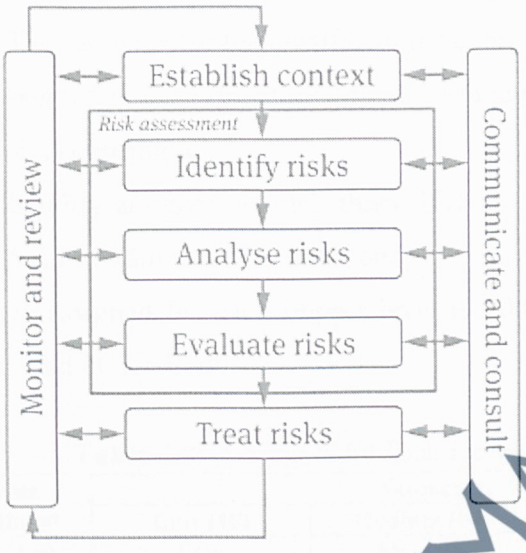
and produce high quality products or services where these are the organizational objectives. While for individuals, well managing risk will keep them away from any harm and loss.

The focus of risk management is the assessment of significant risks and the implementation of suitable risk responses. The objective is to achieve maximum sustainable value from all the activities of the organization. Risk management enhances the understanding of the potential upside and downside of the factors that can affect an organization includes an individual.

Risk management is a process that is underpinned by a set of principles. The risk management process can be presented as a list of coordinated activities. According to AIRMIC (2010), managing risk always based on PDCA cycle or Plan-Do-Check-Action phases. The PDCA cycle was originally formulated by Walter Shewhart in 1930's (n.a., 2012c). This basic cycle also applied by many standards established, which include ISO (ISO 27000 and ISO 9000), NIST and PCI (Pelnekar, 2011).

This PDCA cycle has been applied as in the real processes involved in risk management. According to Lark (2015), this is aligned with the sequence of steps and supports for continuous improvement. The common standard usually proposes the four (4) steps of PDCA as:

i. Plan - designing the risk management framework;
ii. Do – implementing the risk management;
iii. Check - monitoring and review the framework; and
iv. Act - continuous improvement of the framework.



Source: Lark (2015)

Figure 2.10: The ISO 31000:2009 Risk Management Process

As depicted in Figure 2.10, risk assessment is part of risk management which involves the identification of risks followed by their analysis and evaluation or ranking. Recognition and ranking of risks together formed the risk assessment activity. Risk treatment is covered in monitoring phase which risk is treated and a response action is according to its impact level; to accept, avoid, reduce, transfer or share the risk.

2.7.2. Risk Level Matrix

Alotaibi et al. (2017) and Zhang et al. (2011) accentuated in their study that some adverse impacts are impossible to be financially measured. Stoneburner et al. (2002) enhanced that impact such as loss of public confidence, loss of credibility, and damage to an organization's interest cannot be measured in specific units but can be qualified or described in terms of high, medium, and low impacts. Authors Stoneburner et al. (2002) in Risk Management Guide for Information Technology Systems developed this guide to designates and describes the qualitative categories; high, medium, and low impact.

The determination of these risk levels or ratings (high, medium and low) may be subjective. The rationale for this justification can be explained in terms of the probability assigned to each threat likelihood level and a value assigned for each impact level. For example,

i. The probability assigned to each threat likelihood level is 1.0 for High, 0.5 for Medium, 0.1 for Low; and

ii. The value assigned for each impact level is 100 for High, 50 for Medium, and 10 for Low.

Table 2.7: A Sample Of Risk Level Matrix

Threat Likelihood	Impact		
	Low (10)	**Medium (50)**	**High (100)**
High (1.0)	Low 10x1.0=10	Medium 50x1.0=50	High 100x1.0=100
Medium (0.5)	Low 10x0.5=5	Medium 50x0.5=25	High 100x0.5=50
Low (0.1)	Low 10x0.1=1	Medium 50x0.1=5	High 100x0.1=10
Risk Scale: High (51-100); Medium (11-50); Low (1-10)			

Source: Stoneburner et al., (2002)

This risk scale, with its ratings of High, Medium, and Low, represents the degree or level of risk. The same methodology in calculating risk level is proposed by PCI DDS ((SIG), 2012).

2.7.3 Description Of Risk Level

Every level of risk has distinct description according to its magnitude of impacts. Different level of impacts also initiated a different type of action in responding to threat for dissimilar field domains. Usually, high risk reflects catastrophe effect that is difficult to handle, while for the moderate or medium level of risk depict the effect that is lesser than high but still dangerous or could bring severe damage. However, low risk is the level of hazard could be very minimal or nearly secure with no significant damage ((SIG) (2012); Blank & Gallagher (2012); Stoneburner et al. (2002).

2.8 Text Mining

Text mining is part of data mining that extracts the useful information and knowledge hidden in text content, which usually from unstructured text documents. This unstructured text tends to be free-form, non-tabular, dispersed, and not easily retrievable. Hence, it requires deliberate intervention to make sense of it (Nayak et al., 2016). Through techniques such as categorization, text mining is able to be applied in many applications such as entity extraction, sentiment analysis, information retrieval, document classification and clustering (Witten, 2011). This text or content analysis commonly consists of a few processes that include data collection, text pre-processing or also known as pre-treatment, attribute selection that comes with weighting value, the discovery of a pattern and finally the possible prediction that is interpreted by the whole process (Wasilewska, Vijaykumar & Sonawane, 2014).

Text mining is not only used for document organization, spam filtering, hierarchical categorization of web pages and text filtering (Bali & Gore, 2015) and opinion mining (Gasanova, Sergienko, & Semenkin, 2014; Samsudin, Hamda, Puteh, & Nazri, 2013), but also in detecting trend and pattern which include to detect cyber-crime (Kontostathis, Edwards, & Leatherman, 2010).

2.8.1 Pre-processing Of Text

Text pre-processing or also known as pre-treatment of text is intended to represent each document as a feature vector, that is, to separate the text into individual words (Srividhya & Anitha, 2010). Four (4) common pre-processing steps of text classification including tokenization, stop word removal, lowercase conversion, and stemming are commonly considered. Authors in Nayak, Kanive, Chandavekar, & R, 2016; Vijayarani, Ilamathi, & Nithya, 2015; S.Karman & Gurusamy, 2015 and Uysal & Gunal, 2014 agreed that these four (4) processes that construct the pre-processing phase have the following technique description.

i. Tokenization - the process of breaking a stream of text (or documents or sentences) into words, phrases, symbols, or other meaningful elements called tokens. The aim of the tokenization is the exploration

of the words in a sentence. The list of tokens becomes input for further processing such as parsing or text mining.

ii. Stop word removal - Stop-words make the text heavier in terms of space and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text documents are articles, prepositions, and pronouns, which do not give the meaning of the documents and treated as stop words. Stop words are removed from documents because those words are not measured as keywords in text mining applications.

iii. Lowercase conversion or also known as capitalization - Since uppercase or lowercase forms of words are assumed to have no difference, all uppercase characters are usually converted to their lowercase forms prior to the classification.

iv. Stemming or also known as lemmatization - This method is used to identify the root or stem of a word. For example, the words connect, connected, connecting, connections all can be stemmed to the word "connect". The purpose of this method is to remove various suffixes, to reduce the number of words, to have accurately matching stems, to save time and memory space and hence reduce computational cost.

2.8.2 Effects Of Pre-processing

In text mining, pre-processing is not a mandatory process. However, this particular phase has its own effect (either positive or negative) on the accuracy of classifier classification. According to S.Kannan & Gurusamy (2015), pre-processing may reduce indexing or data file size of the text documents, which stops words account about 20 to 30% of total word counts in particular text documents and stemming may reduce indexing size as much as 40 to 50%. Furthermore, this phase may improve the efficiency and effectiveness of the text mining system. However, authors claimed that stop words are not useful for searching or text mining and they may confuse the retrieval system and need to

be removed. Other than that, stemming used for matching the similar words in a text document.

Srividhya & Anitha (2010) proved in their study that pre-processing giving huge impact on performances of classification of the Reuter 21578 dataset. In another study of Gonçalves & Quaresma (2005), removing the stop-words and executed lemmatisation is beneficial for the classification, and in H. Zhang & Wang (2009) it enhances the performance of the Support Vector Machine (SVM) classifier.

However, this is seems resulted in opposite opinion of pre-processing task for spam classification. Almeida, Gomez, & Yamakami, (2011) and Almeida, María, Hidalgo, & Silva (2012) claimed that pre-processing could weaken the value of accuracy for spam classification. While, a study by H. Zhang & Wang (2009) in the same field of distinguishing spam, proved that this phase is extremely important as it has direct relation and effect with the quality of classification outcome. But, surprisingly in a later paper of Almeida, Silva, Santos, & Gomez Hidalgo (2016), authors affirmed that the process of pre-processing does facilitate in increasing the accuracy value.

This research is executing both pre-processing phases and without it to evaluate the effect of pre-treatment in measuring the risk concentration of text spam messages. This is also to verify whether this particular phase is enhancing the classifier accuracy.

2.8.3 Statistical Analysis For Weight Derivation

As articulate that Danger Theory is probably suitable to be applied in this spam risk calculation, it is critical to identify the right feature extraction. Danger Theory that is known as signal processing algorithm indicated that this research needs a method that somehow studies the context of messages (categorical data) to be transformed into input signals (numerical value) for further processed by the Danger Theory algorithm.

Looking at this issue, statistical analysis is the most appropriate method to gain weights for every word and term in short text messages. This analysis that offers numerous term weighting schemes either supervised or unsupervised

method (Patra & Singh, 2013) is able to calculate the weights derivation for input signals in Danger Theory signal processing algorithm requirements.

It is realized that the algorithm from Danger Theory, that is DCA stipulated that the danger value is in between 0 to 1 and the closer the value to 1 indicating that the antigen is most likely dangerous (Greensmith et al., 2009). Considering this critical criterion, term weighting schemes that derived the value in between 0 and 1 and the closer the value of tokenized word to 1 indicate it is more important in spam category and considerably as high risk.

This is also influenced by the dataset collection. The weighting schemes that calculating from the global dataset is always intensified for weight derivation because it is influenced by a larger collection of the dataset where the statistics are derived from (Lintean, Moldovan, Rus, & McNamara, 2010).

2.8.4 Term Weighting Schemes In Spam Classification And Risk Assessment

Each term in a document vector must be associated with a numerical value called weight, which measures the importance of this term and denotes how much this term contributes to the categorization task of the document (Patra & Singh, 2013). In addition to that, Srividhya & Anitha (2010) affirmed that different terms in different document categories have a significantly distinguished level of importance in particular category. The weight is associated with every term as an important indicator and commonly refers the more frequent of terms occurred in a dataset as high ranking keywords for its importance (Arago, Erigieri, Yanoguti, & Paiva, 2016).

Reviews of various feature selection methods are conducted and it is found that Term Frequency (TF) or also known as term strength, Information Gain Ratio (IG Ratio) and Chi-square (CHI^2) is the mostly utilized weighting schemes in text categorization, especially in spam classification for email and short text messages. This is usually produced a high accuracy in the regards field. The references associated with this are tabulated in Table 2.8 as below.

Table 2.8: Term Weighting Schemes

Term Weighting Schemes	References
Term Frequency (TF)	Fernandes, Freire, Fazendeiro, & Inácio (2017); Adewole, Anuar, & Kamsin (2016); Sethi & Bhootna (2014); H. Zhang & Wang (2009)
Information Gain Ratio (IG Ratio)	Trivedi & Dey (2016); Mohamad Mohsin, Hamdan, & Bakar (2015); Ozarkar & Patwardhan (2013); Gansterer & David (2009)
Chi-square (CHI^2)	Waheeb & Ghazali (2017); Trivedi & Dey (2016); Warade et al. (2016); Zareapoor & Seeja (2015); Mohamad Mohsin, Hamdan, & Bakar (2015); Ozarkar & Patwardhan (2013)

It is shown that this term weighting schemes are able to provide positive discrimination on frequent and infrequent terms. Hence, these three (3) weighting schemes turn out to be the benchmark point for this research in applying the same schemes in weight derivation. The weight gained from these schemes is treated as the input signals that then is further calculated for its risk concentration using the developed DCA algorithm model.

2.9 Dataset

Abdulhamid et al. (2017) insisted that accessibility to a requisite dataset constitutes one of the challenges researchers often face in successfully carrying out research on filtering or classifying SMS spam messages. Authors also explored and refined a list of credible research dataset used by researchers for the study in the field that require SMS or short text messages. In addition to that, as described in Oda (2005), a good corpus of the dataset for spam experiment definitely required in order to verify its accuracy detection empirically. The specifications of good corpus include:

i. Should be publicly available – this makes it easier for others to verify results by testing other frameworks using the same corpus. Only then, identification of better framework could be detected.

ii. Contain spam and non-spam – as in spam management, the first phase start with the spam detection, hence the availability of both spam and non-spam messages in a corpus would help much in the experiment as the pre-testing phase. Absolutely as for this research, spam messages are the major requirement.

iii. Messages must be sorted – the messages in a corpus are required to be pre-labelled as spam or non-spam to assist the researchers.

iv. Recent as possible – Over time, techniques used by spammers have changed. Then, having the recent corpus is essential to ensure relevancy and accuracy of the designed framework that match with the current features of spam.

v. Altered as little as possible – it is important to sustain the corpus messages as much as possible as to establish the integrity when run the messages during training and testing phase.

There is a various type of spam and each of it has uniqueness in terms of its identification. SMS messages are usually shorter than email messages and only contain text. There are only 160 characters (which 1 character is 7 bits) or also equal to 140 bytes (which 1 byte is 8 bits); are allowed in a standard text SMS. In spam detection, this fewer words in SMS messages could be a problem in analysing messages because there is less information to work with. Almeida et al. (2011) affirmed that users tend to use acronyms when writing SMS and abbreviations used by SMS users are not standard for a language and based on colloquial of those users communities. Due to these reasons, it could affect the spam filtering accuracy. In text mining, these always referred as noise that would weaken the classifier for spam detection. Even though usually spam is always written in a formal language, other noise can be cleaned up via the process of text pre-treatment or pre-process.

According to Marzuki (2013), SMS linguistic features that commonly required to be considered for many types of study that is not limited to spam detection may consist of the following:

i. the omission of vowel or consonant. For example, nvr is never, appt is appointment

ii. colloquial deviation or standard system deviation. For example, tym is time, gonna is going to

iii. emoticon or textonomatopoeia (spellings which resemble sounds like laughter and cry). For example, ☺ is smile face

iv. rebus abbreviation. For example, l8tr is later, 2mro is tomorrow

v. coded abbreviation. For example, gtg is got to go, asap is as soon as possible

SMS also referred as short text messages. Song, Ye, Du, Huang, & Bie (2014) defined short text message with the main characteristic is no longer than 200 characters. This would include a mobile short message, instant message, Bulletin Board Systems (BBS), online chat record, blog and news comment. The closest structure in term of length with SMS would be a Twitter message that is only 140 characters (Wang, 2012). Even though SMS messages seems to have the same architecture with another type of short text messages, this does not make that the same filters could work with the same efficiency. All these different types of messages might be slightly dissimilar from each other.

In other work, Almeida et al. (2011) and Hidalgo, Bringas, & Sánz (2003) claimed that algorithms and classifiers that successfully in detecting email messages not necessarily are producing the same result for distinguishing SMS spam messages. However, those algorithms can be applied with some modification due to the format structure for SMS and email messages are distinguished from each other.

Since this study is focusing on content features of SMS messages, it usually consists of spam keywords, URL links, monetary value, special characters, emotion symbols and function words. While SMS Service Centre or SMSC commonly applied non-content features for their spam detection that consider message metadata such as length, the number of characters, white spaces, the number of terms, date, time and location.

2.9.1 Type Of Dataset Sources

There are three (3) available sources of the dataset for any research that required SMS messages, both for spam and ham. Authors of this paper elaborated that the sources of the corpus are from public online shared, synthetic or artificial dataset and own collection gathered by researchers. Every type of source has the different objective of the collection.

The largest collection and most utilized dataset for SMS messages study are the online shared for public access, UCI Machine Learning Repository (Almeida & Hidalgo, 2012). Authors who collected and shared this UCI corpus has performed duplicate analysis based on plagiarism detection techniques and

implemented with tool WCopyfind[12]. They also insisted that the results from various experiments by utilizing the same data set could improve the methodology employed and the reliable solution would be remarked (Almeida et al., 2011; Almeida et al., 2012). For example, a different methodology has been applied for SMS spam filtering in Al-Hassan & EL-Alfy (2015) but with the same deployment of UCI corpus. A comparative analysis could be done in identifying which study has contributed a better result in terms of methodology in the world of research. Not limited to spam filtering, the same corpus can be utilized for different objective and the prospect of the research, such as spam clustering or categorization conducted in Delany et al. (2012).

The synthetic or artificial corpus that is generated in study Yoon, Kim, & Huh (2010) is purposely demanding for considering a wide range of application environments, which each of it requires a different level of accuracy and traffic usage.

2.9.2 Available SMS Dataset

The mandatory item for this study is the availability and accessibility to the SMS messages corpus. Without dataset, the developed algorithm is impossible to be tested and implemented, and as a result, could not be verified for its significance. For any type of study that needs the involvement of short text messages, some researchers already have shared the corpus in various sizes and multi languages.

As elaborated in the above paragraph, the largest English corpus for SMS messages are shared publicly at UCI Machine Learning Repository (Almeida & Hidalgo, 2012). Besides this corpus, there is some more available shared English dataset but with the smaller size of the corpus which can be found in Grumbletext, Caroline Tag's, National University of Singapore (NUS) and DIT SMS Spam Dataset (Abdulhamid et al., 2017). There are also shared dataset with the smaller size of the corpus that has been utilized by Narayan & Saxena (2013) in their simulation.

[12] http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/

For dataset in other languages such as Italian, Swedish, German, French, Chinese and Malay can be found in Foozy, Ahmad, & M.A. (2014). This is not limited to SMS messages only but includes another type of short text messages such as opinion messages from various web pages.

2.10 Summary

Many researches have been found for spam filtering is considering content features, but not in numerical value. The novel feature of this research is the weight assigned to the content of the text spam message. Every term in the message content has its own value indicating the importance of that term in spam category and the intensity of its importance also able to be measured.

The final value in determining its risk concentration is derived via equation applied in Danger Theory through signal processing in DCA. Different risk scale range value and weight for signal transformation are deployed to analyse its sensitivity. This suggestion is proof through experiment that will be explained in Chapter 5 for findings and results. Furthermore, the major goal of the classification algorithm is to maximize the predictive accuracy obtained by the model.

It is also proven that every type of spam is unique and various mechanisms for an anti-spam solution are required for a different type of spam. And every mechanism demonstrated a different level of accuracy for its classification or detection rate.