Praise Worthy Prize

International Review on Computers and Software (I.RE.CO.S.), Vol. 9, N. 7
ISSN 1828-6003 *July 2014*

A Framework for SMS Spam and Phishing Detection in Malay Language: a Case Study

Cik Feresa Mohd Foozy, Rabiah Ahmad, Faizal M. A.

Abstract – *Short Message Service (SMS) spam and SMS phishing has been increase nowadays especially in Malay language which is the first language for Malaysia country. Currently, many SMS spam in others language has been proposed, however not yet for Malay language and we are the first to propose these. In addition, this paper also analyst on several frameworks of SMS spam filtering for our SMS spam and phishing detection framework. From the analysis, the chosen framework has been enhanced for Malay SMS spam and phishing. The enhancement has been done on classification phase where our framework proposed dual classification. The classification 1 will classify the SMS into ham and scam SMS. For classification 2, the scam SMS will be classified again into SMS spam and SMS phishing. After dual classifications phase completed, the Malay SMS has been examined using Naïve Bayes and J48 unsupervised Machine Learning techniques. The result shows high accuracy in detecting Malay SMS ham, spam and phishing.*

Keywords: *Detection, Filtering, Phishing, Security, SMS, Spam*

I. Introduction

SMS is one of the alternative communication services nowadays. Since the SMS charges are in reasonably priced per SMS in many countries, these gain interest to many users that don't have mobile data or Internet to use this service as communication tool and sending information such as advertising marketing, spread news and etc. There are tools and software for sale to filter SMS. However, the SMS spam and phishing attack is still increased. Spam and SMS phishing is two different types of attack. Spam message will contain advertising [1] and marketing. However, for phishing attack the message will trick users by announce user is a winner or get free gift. These phishing tactics is for user to response the message. According to [2] the SMS phishing messages can charged fees to mobile device owner silently if the user response to the SMS. Moreover, from the replied SMS phishing, the phishes can get more information about the mobile device such as mobile device version, contact number and etc.

Boodae [3] identify, mobile device users are three times more likely to enter a web-based phishing attack than desktop users and a security report by Lookout [4] found three out of ten Lookout's users interested to clicking on an unsafe link per year by using mobile device. Recently, there are variety of mobile applications have been developed and have been download on mobile device by many user. This is one of the reason phishes revolutionize their attack strategy by embedding malware URL links into the SMS for user to clicks or reply the SMS as accepting the rules and regulations to download or install the mobile applications.

Moreover, if user clicked or replied the trigger SMS, the malware will connect with the mobile device and the mobile device owner has potential to losing money and information privacy. Joe et al. [5] found the unwanted incoming SMS makes the respondent feel the SMS violate their personal privacy and SMiShing usually will influence the SMS recipient to enter their fraud contest, asking money transactions into their bank account, pay their bills, fraud advertising and etc. Since there are many disadvantages of SMiShing attack, few numbers of studies on SMS spam filtering has been proposed.

Traditionally, phishes will send SMiShing attack in a high volume. However, nowadays phishes interested to send SMiShing attack in a small quantity to observe the security level of the mobile device, thus it is possible to detect the SMS phishing based on the SMS quantity and the distribution of sending message pattern.

In addition, SMS language or abbreviation words are frequently used when sending the SMS instead of proper language because the limited text length to write the message. This SMS compose features on mobile phone will make user compress the message by using abbreviation words as long as the recipients read the message. However, too much abbreviation words in SMS will be detected as spam by some of the SMS filtering tools.

In this paper we proposed a framework to detect Malay SMS spam and phishing using classification techniques. Malay SMS corpus, have been collected from the contributors such as web site, friends, family and unknown respondents.

The text collection method is by using transcription and our online form.

This paper is organized as follows. In Section 2, Literature review on SMS phishing detection.

In Section 3, will explained the pre-processing SMS datasets, features selection development and the experiments to identify the accuracy of features selection using data mining tools named WEKA. Section 4, shows the result and findings based on the experiment that has been done and finally is conclusion and future work were given.

II. Related Work

Mobile device phishing attack has been categorized by Dunham et al.[6] into four types such as Bluetooth phishing, Short Message Service (SMS) phishing and Voice over IP Phishing(vishing). Example Bluetooth phishing attack has been discuss by [6], the Bluetooth phishing attack works when user connect to the Wi-Fi hotspot and the attacker can steal the data when the user connects to the Wi-Fi. For vishing, this attack will attack customers' service organization to get the private information of customers.

Mobile web application phishing is similar with desktop web phishing attack. However, the solution for this attack mostly based on client-server method and depends on the mobile architecture itself because mobile device has few limitation such as CPU processing, battery life and memory size, thus the existing solution is by using antivirus or anti-phishing which is purposely develop for lightweight mobile device. Thus, this paper, we will focus on SMS spam and phishing attack solution that has been increased recently.

II.1. SMS Corpus Collection Method

Several methods can be used to collect text corpus.

According to Table I, there is variety of text size that has been collected. Most SMS languages that has been collected are in English SMS and text [7] contributor are from unknown, known contributor and participant.

For unknown SMS contributor are usually SMS taken from software and websites.

For known are from family and friends. Few studies not mention the process of SMS collection clearly.

However, Hard af Segerstad [8], Ogle [9], Fairon and Paumier [10], Choudhury [11], Herring and Zelenkauskaite [12], Bach and Gunnarsson [13] has state the method that has be used to collect SMS.

SMiShing attack has been discussed by [6] and [39]. According to K. Dunham [6], SMiShing is a new tactic to spread malware by adding the URL link in SMS and influence the recipient to click on the URL.

In addition O. Salem et al. [39], identify the SMiShing attack is on the rise attack currently on mobile device and Xu et al. [40] agreed SMS spamming is a serious attack for SMS nowadays. J. W. Yoon et al. [41] and H. Peizhou et al. [42] studies on SMS content-based filtering and Q. Xu et al. [40] proposed SMS filtering on non content-based.

For this paper, the Malay SMS phishing and spam will be classified based on the generic features such as Total words [43], [44], [45] and [46], Number of Character bi-grams[43] and [45], Number of Character tri-grams[43] and [45], Average number of length [44] and [46] and Average number of word [44] and [46].

We also add additional features for this study such as advertisement, contest, urgent SMS, ask money, asked to response SMS, telephone, URL and announce user get free gift or win. These criteria is based on spam and phishing attack. Total features applied in this study are 14 features.

II.2. Malay Language Detection Framework

Malay language has been applied as case study for several research areas. However, there is still lack for information security detection study especially for SMS spam and phishing study which is not yet been examined based on the literature review in SMS spam and phishing study. The example of Malay detection research area such as speech detection[47], language detection[48],[7] and email spam detection [49].

Moreover, speech detection by Lim et al.[47] purposed to produce high quality synthetic speech in Malay language. In addition, for language detection by Tsai et al. [48].

The researchers aimed to reduce the English, Malay and Chinese languages documents redundancy. The framework has three different preprocessing which suit for three different languages and finally will apply in a detection process such as Sentence Segmentation, Sentence Level and Novel Rate computation.

A. Kwee et al. [7], proposed English and Malay language detection and the processes for Malay language detection consist of Language Translation, Stop Word Removal, Word Stemming and Detection.

In email spam detection research area by T. Subramaniam et al. [49], the Malay language has been applied as case study using Bayesian filtering techniques.

The results show this technique successfully classify spam and non-spam of Malay language email with 96% accuracy. The frameworks consist of Collection of emails, Tokenization, Features reduction, Features selection, Training and Testing.

Thus, in this paper, Malay language has been examined as case study for SMS spam and phishing detection. The basic detection process has been applied in this framework and the enhancement of dual classification with two additional new features such as advertisement and contest features has been proposed.

The framework and result will discussed further in the next section.

II.3. SMS Filtering and Detection Framework

Generally, filtering method consist of tokenization, lemmatization and stop word removal, representation and classifier by [50].

TABLE I
LITERATURE ON TEXT MESSAGING CORPUS

References	Text Size	Text Language	Text Contributor	Text Collection Method
Pietrini [14]	500	Italian	15–35 Years Old	Unknown
Schlobinski et al. [15]	1500	Germany	Students	Unknown
Shortis [16]	202	English	1 Male Student, friends and Family	Transcript
Doring [17]	1000	German	200 participants	Unknown
Hard af Sergerstad [18]	1152	Swedish	112 from an anonymous webpage, 252 messages forwarded from volunteers and 788 from family and friends	From webpage, volunteers, family and friends
Kasesniemi and Rautiainen [19]	7800	Finnish	Teenagers (13–18 Years Old)	Transcript
Grinter and Eldridge[20]	477	English	10 Teenagers (15–16 Years Old)	Transcript
Thurlow and Brown [21]	544	English	135 Freshmen	Transcript
Ogle	97	English	Nightclubs	Subscribe SMS Promotion Of Nightclub
Yijue How and Kan [22]	10117	English	2003 Respondents	Transcript
Fairon and Paumier [10]	30000	French	166 University Students	Forward
Choudhury [11]	1000	English	3,200 Contributors	Search The SMS From The Website
Rich Ling [23]	867	Norwegian	Randomly	Transcript
Rettie (2007)	278	English	32 contributors	Unknown
Žic Fuchs and Tuđman Vuković [24]	6000	Croation	University students, family and friends	Unknown
Gibbon and Kul [25]	292	Polish	University students and friends	Unknown
Deumert and Oscar Masinyana [26]	312	English, isiXhosa	22 young adults	Transcript, Forward
Hutchby and Tanna [27]	1250	English	30 young professionals (20-35 years old)	Transcript
Walkowska [28]	1700	Polish	200 contributors	Forward, Software
Herring and Zelenkauskaite [12]	1452	Italian	Audiences of an iTV program	Online SMS archives
Tagg[29]	10628	English	16 family and friends	Transcript
Elvis [30]	600	English, French, etc.	72 university students and lecturers	Forward
Barasa [31]	2730	English, Kiswahili, etc.	84 university students and 37 young professionals	Forward
Bach and Gunnarsson [13]	3152	Swedish, English, etc.	11 contributors	Software
Bodomo [32]	853	English, Chinese	87 youngsters	Transcript
Liu and Wang [33]	85870	Chinese	Real volunteers	Unknown
Sotillo [34]	6629	English	59 participants	Software
Dürscheid and Stark [35]	23987	Germany, French, etc	2,627 volunteers	Forward
Lexander [36]	496	French	15 young people	Transcript
Elizondo [37]	357	English	12 volunteers	Transcript
Chen and Kan [38]	71000	English and Mandarin Chinese.	University Students	Unknown

However, there are multiple frameworks for spam filtering and detection such as below:

- *Tools*

Cormack et al. [45] on SMS filtering using 5 types of spam filter tools to filter the SMS. Before applied tools on SMS, there are pre-processing data that has been done to come out with four (4) main features. Moreover, to detect SMS protocol in real time Rafique et al. [51] applied Hidden Markov Model (MHH) which the architecture consists of sniffer, feature extraction, classifier and rules decision. Que and Farooq [52] also apply MHH on byte level distribution of SMS that have four processes such as identify suitable representation of an SMS, build spam models, classification and determined the spam. In addition, SMS-Watchdog SMS detection scheme by Yan et al. [53] has three processes of monitoring, anomaly detection and alert handling using SMS services.

- *Content-Based Filtering*

J. W. Yoon, et al. [41] proposed hybrid framework that implement content-based technique with challenge-response scheme. The SMS classified into ham, spam and uncertain, then the challenge response will classify uncertain into ham and spam by matching the sender

response. Gòmez Hidalgo et al. [54] have proposed content-based SMS filtering for English and Spanish SMS spam using Bayesian filtering that consist of preprocessing, feature selection and learning.

- *Machine Learning*

Xiang et al. [55] proposed Support Vector Machine technique to filter the mobile spam. Moreover, Cai et al. [56] improved the spam filter using traditional balanced Winnow algorithm which applied pre-processor, feature selection, texts representation and winnow algorithm module. An independent mobile device filtering by Taufiq Nuruzzaman et al. [44] applied several processes in their SMS independent spam filtering such as data set and running environment, feature extraction, vector creation and filtering process of Naïve Bayes or SVM and update filtering system. Yadav [46], [57] had three process in the their SMS filtering such as Bayesian filtering algorithm, mobile application and synchronization service on server.

- *Matching Pattern*

Wu et al. [58] has proposed SMS filtering flow such as SMS screening, bayesian learning, keyword SMS and Pinyin Fuzzed keyword matching.

Moreover, the Chinese SMS filtering by Jie et al. [59] has pre-processing, features selection, modeling, and classifier. In addition, Najadat et al. [60], frameworks involve of three processes of data collection, pre-processing, text mining, testing, evaluation metrics and implementation.

- *Artificial Immune System*

T. M. Mahmoud and A. M. Mahfouz [61] applied artificial immune system method filter SMS spam that contain analysis engine, tokenize word, stop word, dataset, training and AIS engine. Charninda et al. [62] proposed a hybrid solution of neural network and Bayesian filtering where the SMS filtering process are sender identification module, spam folder, SMS content extractor, tokenizer, Bayesian filter, categorization, training and inbox.

- *Cryptography*

In Cryptography area, Saxena[63] proposed a secure SMS protocol for SMS transmission and a cryptographic algorithm in the SIM card. The processes of framework are request to send SMS and authenticate sending SMS.

In addition, Pereira et al.[64] also proposed a lightweight cryptography algorithm to mitigate the SMS security issues, protocols, providing encryption, authentication and signature services. In addition, Choi [65] applied *Common Public Key Cryptography* technique for SMS communication effeciency which containt of initialization for authenticate, encrypt, or decrypt and communication phases for sending SMS.

As summary, the basic architecture frameworks are data collection, pre-processing, features selection, training and testing.

Which is mostly has been applied in SMS studies.

These explained the SMS spam filtering study already applied various filtering and detection techniques with different SMS language except for Malay SMS language.

Additional, one of the different between frameworks is the technique applied but yet the result is still good.

Varieties of framework presented for spam filtering and detection. However, for SMS spam and phishing attack not yet available. Thus, the proposed SMS spam and phishing framework will do some enhancement on the generic framework of SMS spam filtering by [50].

The enhancement framework will have dual classification for SMS Malay language.

The reason to have dual classification is to identify the SMS collection has been classified correctly. After the first classification done, the second classification process will classify the scam SMS into spam and phishing. The framework will discuss further in next section.

III. Methodology

This section, explain about Malay SMS spam and phishing detection framework development. The SMS spam and detection will focus on features based on previous studies. As mention before, this study is the first to collect Malay SMS for detecting spam and phishing.

Thus, there are no SMS spam and phishing datasets available in Malay Language. SMS spam and phishing datasets need to be prepared for this study.

For datasets preparation, a collection of Malay SMS has been done from website, unknown respondents, friends and family. The proposed framework is based on Guzella and Caminhas [50] which have four (4) main steps in filtering spam message such as tokenization, lemmatization, representation and classifier.

For this framework, four main steps will be applied.

However, additional classifier will be added in this framework which called dual classifier in Malay SMS spam and phishing detection framework. .

The reason we need dual classifier compare to a single classification process because we collect SMS ham and scam SMS from website, friend, family and unknown respondents. The respondents usually have basic knowledge about SMS spam and phishing and some doesn't know anything about these attacks.

After Malay SMS collection have done SMS ham and SMS scam will be tokenizing, lemmatizing and stop word removal, representation and classification 1.

After get the result from classification 1, second classification process will be proceed to classified again SMS scam into SMS spam and phishing. The similar method has been applied by J. W. Yoon, et al. [41] to classify uncertain SMS into spam and ham class. Figure 1 and the process below listed the process of Malay SMS spam and phishing datasets and detection development:

i. Collect SMS ham and scam SMS from website, friend, family and respondents and do first classification.
ii. Tokenization.

iii. Lemmatization as remove redundancy and noise.
iv. Representation Strings into nominal datasets.
v. Features Selection.
vi. Second classification to Scam SMS into spam and phishing.
vii. Examine the result Malay SMS datasets using Naïve Bayes and J48 Technique.

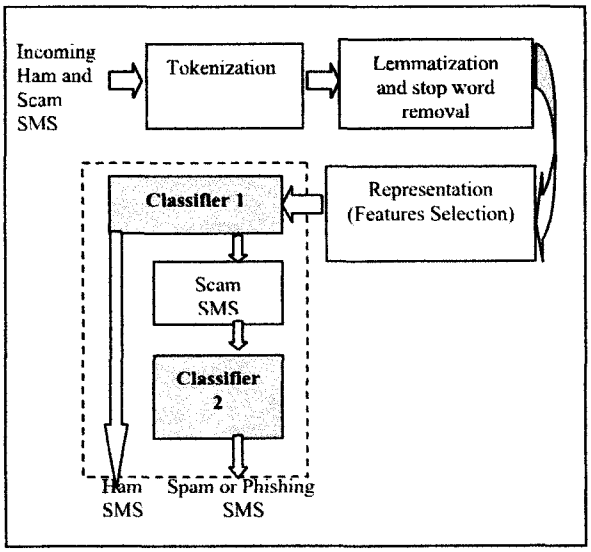

Fig. 1. SMS Spam and Phishing Detection in Malay Language Framework

III.1. Malay SMS Corpus Collection Method

As preliminary study in collecting Malay SMS corpus, Malay SMS has been collected using methods in Table I.

The SMS collection methods are from website, personal SMS forwarding, transcriptions and online form. The SMS contributors are from respondent, website, family and friends. After the SMS collections are done, all SMS are transcript into Microsoft Office Excel 2007 for tokenization, lemmatization and representation process.

III.2. Tokenization

Tokenization is a process to divide the sentence into word. The purpose tokenizations have been done for calculating the word for features selection and classification process. Fig. 2 is an example of the SMS tokenization. There are 179 SMS has been tokenize and total word after tokenization are 21694.

Fig. 2. After SMS Tokenization Process

III.3. Lemmatization

Lemmatization is a process to group the same meaning words.

However, SMS usually will contain many abbreviation words. It is difficult to group similar meaning for variety of words such as in Malay SMS abbreviation, the word Thank You can be typed as TQ, thank Q, thanks or tengkiu. Thus, for this study, all words in this SMS will be calculated the occurrences and will be identified as different words.

The calculation of SMS word occurrences process are done by using JAVA programming to identified the unique words in these Malay SMS collection. There are 802 words in SMS after lemmatization.

III.4. Features Selection

SMS representation in this paper is applying the features based on the previous studies. The features are Total words, Number of Character bi-grams, Number of Character tri-grams, Average number of length, and Average number of word.

For this study, additional features based on the spam and phishing characteristics also included such as Advertisement or announcement, Contest, Malicious URL, Telephone Number, Winning or Free gift, SMS ask help to get money and SMS ask to respond or subscribe services.

III.5. Classification

There are two classifications processes proposed in this framework. The raw data collection has been classified into SMS ham and SMS scam. After classification process 1, the classification result shows high accuracy. After that, the second classification process, classify the SMS scam into SMS spam and phishing. The reason dual classifications are done because this is the first study to proposed framework in detecting SMS spam and phishing. Thus, to ensure good result in classification accuracy, this dual classification has been proposed and the results will be discussed in the next section.

IV. Analysis and Findings

An experiment has been done to examined 179 of SMS ham, spam and phishing class using WEKA a data mining tools to test the classified accuracy, true positive, true false on Malay spam and phishing corpus using Naïve Bayes and J48. The reason these technique has been applied to tested the classification accuracy rate because these techniques is one of the well known supervised method in machine learning techniques.

Table II is a classification result between 41 SMS ham and 82 SMS scam. The result shows Naïve Bayes and J48 is 100%. Table III is a classification is result for Scam SMS that has been classified into 41 SMS phishing and 41 SMS spam Malay SMS. The result also shows Naïve Bayes and J48 get 100 %. The final result for ternary classification of 41 SMS ham, 41 SMS phishing and 41 SMS spam show 100% accuracy.

TABLE II
BINARY CLASSIFICATION 1 RESULT FOR MALAY SMS HAM AND SCAM

Parameter	Naïve Bayes		J48	
	Ham	Scam	Ham	Scam
True Positive	1	1	1	1
False Positive	0	0	0	0
Correctly Classified	**100 %**		**100%**	
Incorrectly Classified	**0 %**		**0%**	

TABLE III
BINARY CLASSIFICATION 2 RESULT FOR MALAY SMS SPAM AND PHISHING

Parameter	Naïve Bayes		J48	
	Phishing	Spam	Phishing	Spam
True Positive	1	1	1	1
False Positive	0	0	0	0
Correctly Classified	100 %		100 %	
Incorrectly Classified	0 %		0 %	

TABLE IV
TERNARY CLASSIFICATION RESULT FOR MALAY SMS HAM, SPAM AND PHISHING

Parameter	Naïve Bayes			J48		
	Ham	Phishing	Spam	Ham	Phishing	Spam
True Positive	1	1	1	1	1	1
False Positive	0	0	0	0	0	0
Correctly Classified		100 %			100 %	
Incorrectly Classified		0 %			0 %	

The higher percentage (%) of correctly classified parameter is better. Meanwhile the True Positive is 1 showing that the datasets is classified correctly and false positive is 0 means that none of SMS are in wrong class.

V. Conclusion

SMS phishing is still arising. However, only SMS spam corpus available on the Internet. Based on the studies on spam and phishing in information security area, spam and phishing attack have different definition and understanding.

Thus, we initiate to collect SMS phishing in Malay language as alternative mitigation to overcome SMS spam and phishing in Malay language.

Based on the result, we prove that this corpus, successfully been classified into three classes such as SMS ham, spam and phishing using dual classification techniques.

As conclusion, by applying classification techniques using machine learning technique also can detect and filter scam SMS.

There are many techniques can be applied for this research area and this technique is one of the established technique for SMS filtering and detection since processing times by using Naïve Bayes and J48 only takes only 0 seconds to classify and get 100% accuracy.

Acknowledgements

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM), Universiti Tun Hussein Onn Malaysia(UTHM) and Ministry of Higher Education Malaysia for supporting this research.

This research is funded by MOHE under Long Research Grant Scheme LRGS/2011/FTMK/TK01/1R00002.

References

[1] S. S. Chandran and S. Murugappan, "Spam detection and elimination of messages from twitter," *International Review on Computers and Software*, vol. 8, pp. 2438-2443, 2013.
[2] F-Secure, "Mobile Threat Report Q3 2012," F-Secure Labs2012.
[3] M. Boodae, "Mobile Users Three Times More Vulnerable to Phishing Attacks," in *Trusteer* vol. 2012, ed, 2011.
[4] I. Lookout, "Lookout Mobile Threat Report August 2011," 2011.
[5] I. Joe and H. Shim, "An SMS Spam Filtering System Using Support Vector Machine," in *Future Generation Information Technology*, vol. 6485, T.-h. Kim, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 577-584.
[6] K. Dunham, "Chapter 6 - Phishing, SMishing, and Vishing," in *Mobile Malware Attacks and Defense*, D. Ken, Ed., ed Boston: Syngress, 2009, pp. 125-196.
[7] A. Kwee, *et al.*, "Sentence-Level Novelty Detection in English and Malay," in *Advances in Knowledge Discovery and Data Mining*, vol. 5476, T. Theeramunkong, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2009, pp. 40-51.
[8] Y. Hård af Segerstad, *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*: University of Gothenburg, 2002.
[9] T. Ogle, "Creative Uses of Information Extracted from SMS Messages," Undergraduate, Computer Science, The University of Sheffield, 2005.
[10] Cédrick Fairon and S. Paumier, "A Translated Corpus of 30,000 French SMS," in *In Proceedings of Language Resources and Evaluation.*, 2006.
[11] M. Choudhury, *et al.*, "Investigation and modeling of the structure of texting language," *Int. J. Doc. Anal. Recognit.*, vol. 10, pp. 157-174, 2007.
[12] S. C. Herring and A. Zelenkauskaite, "Symbolic capital in a virtual heterosexual market abbreviation and insertion in Italian iTV SMS," *Written Communication*, vol. 26, pp. 5-31, 2009.
[13] C. Bach and J. Gunnarsson, "Extraction of trends is SMS text," *Master's thesis, Lund University*, 2010.
[14] D. Pietrini, "X'6:-(?": The sms and the triumph of informality and ludic writing," *Italienisch*, vol. 46, pp. 92-101, 2001.
[15] P. Schlobinski, *et al.*, "Simsen. Eine Pilotstudie zu sprachlichen und kommunikativen Aspekten in der SMS-Kommunikation," *Networx 22. Online-Publikationen zum Thema Sprache und Kommunikation im Internet*, 2001.
[16] T. Shortis, "'New Literacies' and Emerging Forms: Text Messaging on Mobile Phones," presented at the International Literacy and Research Network Conference on Learning., 2001.
[17] N. Doring, "1 bread, sausage, 5 bags of apples I.L.Y" - communicative functions of text messages (SMS)," *Zeitschrift für Medienpsychologie 3*, 2002.
[18] Y. Hard af Sergerstad, *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*: University of Gothenburg, 2002.
[19] E.-L. Kasesniemi and P. Rautiainen, "Mobile culture of children and teenagers in Finland," in *Perpetual contact*, ed: Cambridge University Press, 2002, pp. 170-192.
[20] R. Grinter and M. Eldridge, "Wan2tlk?: everyday text messaging," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA, 2003.
[21] C. a. A. B. Thurlow, "Generation Txt? The sociolinguistics of young people's text messaging," *Discourse Analysis Online 1(1), 30.*, 2003.
[22] Yijue How and M.-Y. Kan, "Optimizing Predictive Text Entry for Short Message Service on Mobile Phones," presented at the In Proceedings of HCII, 2005.
[23] Rich Ling and N. S. Baron, "Text Messaging and IM: Linguistic Comparison of American College Data," 2007.

[24] M. Žic Fuchs and N. Tuđman Vuković, "Communication technologies and their influence on language: Reshuffling tenses in Croatian SMS text messaging," *Jezikoslovlje*, pp. 109-122, 2008.
[25] D. Gibbon and M. Kul, "Economy Strategies in Restricted Communication Channels. A study of Polish short text messages," 2008.
[26] A. Deumert and S. Oscar Masinyana, "Mobile language choices The use of English and isiXhosa in text messages (SMS) Evidence from a bilingual South African sample," *English World-Wide*, vol. 29, pp. 117-147, 2008.
[27] I. Hutchby and V. Tanna, "Aspects of sequential organization in text message exchange," *Discourse & Communication*, vol. 2, pp. 143-164, 2008.
[28] J. Walkowska, "Gathering and Analysis of a Corpus of Polish SMS Dialogues," *Challenging Problems of Science. Computer Science. Recent Advances in Intelligent Information Systems*, pp. 145-157, 2009.
[29] C. Tagg, "A Corpus Linguistics Study of SMS Text Messaging," Doctor of Philosophy, Department of English, The University of Birmingham, Birmingham, 2009.
[30] F. W. Elvis, "The sociolinguistics of mobile phone sms usage in cameroon and nigeria," *The International Journal of Language Society and Culture*, vol. 28, pp. 25-40, 2009.
[31] S. N. Barasa, *Language, mobile phones and internet: a study of SMS texting, email, IM and SNS chats in computer mediated communication (CMC) in Kenya*, 2010.
[32] A. B. Bodomo, "The Grammar of Mobile Phone Written Language," *Chapter*, vol. 7, pp. 110-198, 2010.
[33] W. Liu and T. Wang, "Index-based online text classification for sms spam filtering," *Journal of Computers*, vol. 5, pp. 844-851, 2010.
[34] S. Sotillo, "SMS Texting Practices and Communicative Intention," *Chapter*, vol. 16, pp. 252-265, 2010.
[35] C. Dürscheid and E. Stark, "SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland," *Digital Discourse: Language in the New Media: Language in the New Media*, p. 299, 2011.
[36] K. V. Lexander, "Names U ma puce: multilingual texting in Senegal," Working paper2011.
[37] J. Elizondo, "Not 2 Cryptic 2 DCode: Paralinguistic Restitution, Deletion, and Nonstandard Orthography in Text Messages," Ph. D. thesis, Swarthmore College, 2011.
[38] T. Chen and M.-Y. Kan, "Creating a live, public short message service corpus: the NUS SMS corpus," *Language Resources and Evaluation*, vol. 47, pp. 299-335, 2013/06/01 2013.
[39] O. Salem, *et al.*, "Awareness Program and AI based Tool to Reduce Risk of Phishing Attacks," in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, 2010, pp. 1418-1423.
[40] Q. Xu, *et al.*, "SMS Spam Detection using Content-less Features," *Intelligent Systems, IEEE*, vol. PP, pp. 1-1, 2012.
[41] J. W. Yoon, *et al.*, "Hybrid spam filtering for mobile communication," *Computers & Security*, vol. 29, pp. 446-459, 2010.
[42] H. Peizhou, *et al.*, "A Novel Method for Filtering Group Sending Short Message Spam," in *Convergence and Hybrid Information Technology, 2008. ICHIT '08. International Conference on*, 2008, pp. 60-65.
[43] G. V. Cormack, *et al.*, "Content based SMS spam filtering," presented at the Proceedings of the 2006 ACM symposium on Document engineering, Amsterdam, The Netherlands, 2006.
[44] M. Taufiq Nuruzzaman, *et al.*, "Simple SMS spam filtering on independent mobile phone," *Security and Communication Networks*, vol. 5, pp. 1209-1220, 2012.
[45] G. V. Cormack, *et al.*, "Feature engineering for mobile (SMS) spam filtering," presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007.
[46] K. Yadav, *et al.*, "SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering," presented at the Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, Phoenix, Arizona, 2011.
[47] Y. C. Lim, *et al.*, "Application of Genetic Algorithm in unit selection for Malay speech synthesis system," *Expert Systems with Applications*, vol. 39, pp. 5376-5383, 2012.
[48] F. S. Tsai, *et al.*, "Multilingual novelty detection," *Expert Systems with Applications*, vol. 38, pp. 652-658, 2011.
[49] T. Subramaniam, *et al.*, "Naïve Bayesian Anti-spam Filtering Technique for Malay Language."
[50] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, pp. 10206-10222, 2009.
[51] M. Z. Rafique, *et al.*, "Application of evolutionary algorithms in detecting SMS spam at access layer," presented at the Proceedings of the 13th annual conference on Genetic and evolutionary computation, Dublin, Ireland, 2011.
[52] M. Z. R. que and M. Farooq, "SMS Spam Detection By Operating On Byte-Level Distributions Using Hidden Markov Models (HMMS)," presented at the Virus Bulletin Conference September 2010, 2010.
[53] G. Yan, *et al.*, "SMS-Watchdog: Profiling Social Behaviors of SMS Users for Anomaly Detection
Recent Advances in Intrusion Detection." vol. 5758, E. Kirda, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2009, pp. 202-223.
[54] J. M. G. Hidalgo, *et al.*, "Content based SMS spam filtering," presented at the Proceedings of the 2006 ACM symposium on Document engineering, Amsterdam, The Netherlands, 2006.
[55] Y. Xiang, *et al.*, "Filtering mobile spam by support vector machine " presented at the Conference on Computer Sciences, Software Engineering, Information Technology, E-Business and Applications (3rd: 2004 : Cairo, Egypt), Cairo, Egypt, 2004.
[56] C. Jie, *et al.*, "Spam Filter for Short Messages Using Winnow," in *Advanced Language Processing and Web Information Technology, 2008. ALPIT '08. International Conference on*, 2008, pp. 454-459.
[57] K. Yadav, *et al.*, "Take Control of Your SMSes: Designing an Usable Spam SMS Filtering System," in *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, 2012, pp. 352-355.
[58] W. Ningning, *et al.*, "Real-time monitoring and filtering system for mobile SMS," in *Industrial Electronics and Applications, 2008. ICIEA 2008. 3rd IEEE Conference on*, 2008, pp. 1319-1324.
[59] J. Huang, *et al.*, "A Bayesian Approach for Text Filter on 3G Network," in *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, 2010, pp. 1-5.
[60] H. Najadat, *et al.*, "Mobile SMS Spam Filtering based on Mixing Classifiers."
[61] T. M. Mahmoud and A. M. Mahfouz, "SMS Spam Filtering Technique Based on Artificial Immune System," *IJCSI International Journal of Computer Science Issues*, vol. 9, 2012.
[62] T. Charninda, *et al.*, "Content based hybrid sms spam filtering system," 2014.
[63] N. Saxena and N. S. Chaudhari, "SecureSMS: A secure SMS protocol for VAS and other applications," *Journal of Systems and Software*, vol. 90, pp. 138-150, 2014.
[64] G. C. C. F. Pereira, *et al.*, "SMSCrypto: A lightweight cryptographic framework for secure SMS transmission," *Journal of Systems and Software*, vol. 86, pp. 698-706, 2013.
[65] J. Choi and H. Kim, "A Novel Approach for SMS security," *International Journal of Security & Its Applications*, vol. 6, 2012.

Authors' information

Cik Feresa Mohd Foozy is currently working with Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia. Feresa holds a Master's degree in Computer Science(Information Security) from Universiti Teknologi Malaysia, Malaysia and a Bachelor's degree in Information Technology and Multimedia from Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia. She is currently pursuing her PhD at the Universiti Teknikal Malaysia Melaka, Malaysia.