Indonesian Journal of Electrical Engineering and Computer Science
Vol. 15, No. 2, August 2019, pp. 1046~1053
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v15.i2.pp1046-1053

Video spam comment features selection using machine learning techniques

Nabilah Alias, Cik Feresa Mohd Foozy, Sofia Najwa Ramli
Applied Computing Technology (ACT), Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received Sep 12, 2018
Revised Jan 20, 2019
Accepted Mar 2, 2019

Keywords:

Video spam comment
Machine learning
Feature selection

ABSTRACT

Nowadays, social media (e.g., YouTube and Facebook) provides connection and interaction between people by posting comments or videos. In fact, comments are a part of contents in a website that can attract spammer to spreading phishing, malware or advertising. Due to existing malicious users that can spread malware or phishing in the comments, this work proposes a technique used for video sharing spam comments feature detection. The first phase of the methodology used in this work is dataset collection. For this experiment, a dataset from UCI Machine Learning repository is used. In the next phase, the development of framework and experimentation. The dataset will be pre-processed using tokenization and lemmatization process. After that, the features to detect spam is selected and the experiments for classification were performed by using six classifiers which are Random Tree, Random Forest, Naïve Bayes, KStar, Decision Table, and Decision Stump. The result shows the highest accuracy is 90.57% and the lowest was 58.86%.

Corresponding Author:

Naqliyah Zainuddin,
CyberSecurity Malaysia, Level 4 Block C,
Bangunan MINES Waterfront Business Park, No. 3 Jalan Tasek,
43300 Seri Kembangan, Selangor, Malaysia.
Email: naqliyah@cybersecurity.my

1. INTRODUCTION

At present, worldwide broadband distribution has increased the number of Internet users. With faster connections, hosting and video sharing services are becoming popular among users [1]. The availability of resources over the Internet and broadband connection enables the emergence of sophisticated new platforms. In this way, YouTube is a one well-known video content publishing platform with social networking features, such as support for posting text comments to provide interactions between producers (channel owners) and viewers [2].

Recently, YouTube has used monetization systems to reward producers, stimulating them to produce high quality original content and increase the amount of visualization. After the use of this system, the platform is flooded with unwanted content, typically low quality information known as spam. Spam is the use of an electronic messaging system to send unsolicited messages, especially advertisements, as well as repeat messages on the same website. For social spam, it can be done in many ways, including mass messaging, cruelty, humiliation, hate speech, malicious links, fake reviews, fake hints, and personal information [3].

Indeed, it is a problem that could become critical. It caused the user disable comments on their videos because the most of comments are spam. Until now, the research to detect the spam YouTube comment using machine learning technique is still lacking.

To overcome the problems that appear, this paper proposed technique used for video sharing spam comment feature detection. This works evaluate the performance of spam comment feature detection using accuracy.

2. RELATED WORK

Spam is related to low quality of information and consists of undesired content [1]. Usually, spam found in texts, video and images [4-5]. Most of spam is used to manipulate internet user to obtain personal information such as phishing and malware. Spam also used to make commercial advertising [1]. For spam message, it usually works by flooding the Internet with the same message in order to force user to receive it. Besides, video spam is a low quality content of the video that publish on YouTube by malicious users [6]. There are many researchers related to spam in the existing study such as blog spam [7], web spam [8], twitter spam [9], email spam [10], YouTube spam [1] and SMS spam [11].

Ham is a message that is not Spam. In other words, "non-spam", or "good message" [12]. It should be considered a high quality of information and meaningful words [13].

Figure 1 shows example of spam and ham comments posted on YouTube.

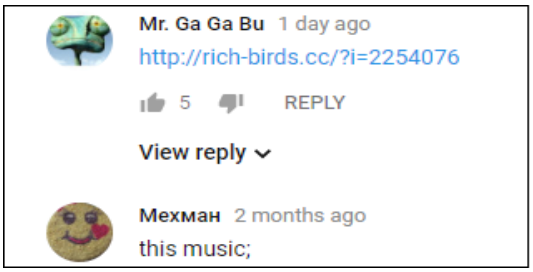

Figure 1. Example of spam and ham comments posted on YouTube

2.1. Pre-Processing

In the pre-processing step, the features are firstly extracted. The subject field contains the data that need to be pre-processing [7]. Therefore, there are a few steps in this phase, which are tokenization, stop words and lemmatization. These processes were doing to remove noise, redundant and also words that common English use that will affect the detection phase [13]. Most of the existing research doing pre-processing process before continuing to the next process. Table 1 shows the spam detection steps and the process used by researchers.

Based on Table 1, the research papers for Tubespam: Comment Spam Filtering on YouTube and Combating Comment Spam with Machine Learning Approaches is using four techniques to detect spam which is Pre-processing, Features Extraction, Classification and Evaluation.

Feature extraction is the process of identifying features or type of information contained within the documents. After these features are extracted, then only the machine learning algorithms can find the target concept descriptions of categories.

The next paper which is Towards Filtering of SMS spam messages using Machine Learning Based Technique is used six techniques, namely Pre-processing, Feature Selection, Classifier Training, Classifier Testing, Classification Result and Performance evaluation.

For research about Statistical Twitter Spam Detection Demystified-Performance, Stability and Scalability are used four techniques which are Data Collection, Feature Selection, Classification and Evaluation.

In paper KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube used only three techniques to detect spam which is Data Collection, Features Selection and Classification.

Next is the research about Detecting Video Spammers in YouTube Social Media, the researcher used four techniques to detect video spammers namely Data Collection, Data Pre-processing, Feature Construction and Classification.

Lastly, research paper with title Data Mining Based spam Detection System for YouTube spam using three techniques for detecting spam which is Data Collection, Classification and Evaluation.

Table 1. Steps and Process for Spam Detection

Author	Title	Detection Technique
[1]	Tubespam: Comment Spam Filtering on YouTube	1. Pre-processing 2. Features Extraction 3. Classification 4. Evaluation
[4]	Combating Comment Spam with Machine Learning Approaches	1. Pre-processing 2. Feature Extraction 3. Classification 4. Evaluation
[13]	Towards Filtering of SMS spam messages using Machine Learning Based Technique	1. Pre-processing 2. Feature Selection 3. Classifier Training 4. Classifier Testing 5. Classification Result 6. Performance evaluation
[9]	Statistical Twitter Spam Detection Demystified-Performance, Stability and Scalability	1. Data Collection 2. Feature Selection 3. Classification 4. Evaluation
[2]	KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube	1. Data Collection 2. Features Selection 3. Classification
[14]	Detecting Video Spammers in YouTube Social Media	1. Data Collection 2. Data Pre-processing 3. Feature Construction 4. Classification
[6]	Data Mining Based spam Detection System for YouTube spam	1. Data Collection 2. Classification 3. Evaluation

2.2. Feature Selection

Feature selection is known as attribute selection, variable selection or variable subset selection. It is the process of selecting a variable for use in model construction. Feature selection techniques are used for four reasons which are:

a. Simplification of models to make it easier to interpret by researchers.
b. Shorter training times.
c. Avoiding the curse of dimensionality.
d. Enhanced generalization by reducing over fitting.

Feature selection is a very important task for the text spam filtering. Selected features should be correlated to the message type such that accuracy for detection of spam message can be increased [11]. Spam and ham messages can be differentiated using various features. Table 2 presents the selected features used to detect spam.

Table 2. Features Selection used in Existing Projects

Features	Author					
Bag-of-words	[1]	[4]	[13]	[9]	[2]	[14]
Post-comment similarity	✓				✓	
Inter-comment similarity		✓		✓		
Interval between post and comment		✓				
Number of words in the comment		✓				
Number of sentences in the comment		✓	✓	✓		
Comment length		✓	✓			
Phone information		✓	✓	✓		
E-mail information		✓	✓			
URL link		✓				
Black word list		✓	✓	✓		
Stop words ratio		✓				
Presence of symbol		✓				
Presence of dots			✓			
Presence of emotions			✓			
Lower-case words			✓			
Uppercase words			✓			
Keyword specific			✓			
Number of digits			✓			
Channel age			✓	✓		
The channel average upload						✓

Based on Table 2, the most features that have been used to detect spam are bag-of-words, post comments similarity, number of words in the comment, number of sentences in comment, comment length, phone information, URL link and number of digits.

2.3. Classification of Techniques

Various techniques used in experiment to evaluate the performance of spam detection. Initially, feature selection is performed and then extracts the features. After extraction, classification of techniques used to get evaluation performance, such as Decision Trees (DTs), Naïve Bayes and so on [13]. Classification techniques are used to detect the accuracy of spam itself. This technique has worked using tools such as WEKA and Rapid Miner. Table 3 shows the machine learning techniques used for six papers.

Table 3. Machine Learning Techniques

Author	Title	Detection Technique
[1]	Tubespam: Comment Spam Filtering on YouTube	1. Decision trees (CART) 2. K -nearest neighbors (k -NN) 3. Logistic regression (LR) 4. Bernoulli Naïve Bayes (NB-B) 5. Gaussian Naïve Bayes (NB-G) 6. Multinomial Naïve Bayes (NB-M) 7. Random Forest (RF) 8. Support vector machines with linear kernel (SVM-L) 9. Support vector machines with polynomial kernel (SVM-P) 10. Support vector machines with a Gaussian kernel (SVM-R)
[4]	Combating Comment Spam with Machine Learning Approaches	1. J48 (C4.5 Algorithm) 2. Random Forest (RFT) 3. Decision Tree 4. SVM 5. Multilayer Neural Network
[13]	Towards Filtering of SMS spam messages using Machine Learning Based Technique	1. Naïve Bayes 2. Logistic Regression 3. J48 4. Decision Table 5. Random Forest
[9]	Statistical Twitter Spam Detection Demystified-Performance, Stability and Scalability	1. K -nearest neighbor 2. Weight K -nearest neighbor 3. Naïve Bayes 4. Random Forest 5. C5.0 6. Boosted Logistic Regression 7. Stochastic Gradient Boosting Machine 8. Neural Network
[2]	KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube	1. Random Forest 2. K-nearest Neighbor 3. Decision Tree
[14]	Detecting Video Spammers in YouTube Social Media	1. Functional Tree 2. J48 3. Random Forest 4. Bayes Network 5. Naïve Bayesian

Based on Table 3, the first author with research about Comment Spam Filtering in YouTube used ten comparison of classification algorithm which are Decision trees (CART), K -nearest neighbors (k -NN), Logistic regression (LR), Bernoulli Naïve Bayes (NB-B), Gaussian Naïve Bayes (NB-G), Multinomial Naïve Bayes (NB-M), Random Forest (RF), Support vector machines with linear kernel (SVM-L), Support vector machines with polynomial kernel (SVM-P) and Support vector machines with Gaussian kernel (SVM-R).

Second author with research about Combating Comment Spam with Machine Learning Approaches used five comparisons of classification algorithm which are J48 (C4.5 Algorithm), Random Forest (RFT), Decision Tree, SVM and Multilayer Neural Network.

Next, for third research which is Towards Filtering of SMS spam messages using Machine Learning Based Technique compared five classification algorithms, namely Naïve Bayes, Logistic Regression, J48, Decision Table and Random Forest.

The fourth research which is Statistical Twitter Spam Detection Demystified-Performance, Stability and Scalability compared eight classification algorithms which are K -nearest neighbor, Weight K -nearest neighbor, Naïve Bayes, Random Forest, C5.0, Boosted Logistic Regression, Stochastic Gradient Boosting Machine and Neural Network.

The fifth research with title KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube have been compared three classification algorithm which is Random Forest, K-nearest Neighbor and Decision Tree.

And the last is Detecting Video Spammers in YouTube Social Media used five classification algorithms to be compared to get high accuracy, namely Functional Tree, J48, Random Forest, Bayes Network and Naïve Bayesian.

The most classification algorithms that have been used by existing research are Naïve Bayesian, Random Forest, Decision Tree and K –nearest Neighbor.

Table 4. Comparison Table in Detection of Spam

Author	[1]	[13]	[9]	[2]	[14]
Year	2015	2017	2017	2016	2017
Accuracy	Above 90%	Above 90%	Above 90%	Above 90%	Above 90%
Algorithm	RF, NB-B	Random Forest	Random Forest,C5.0	Random Forest, K-Nearest Neighbor	Bayes Network, Naïve Bayesian
Type of spam	YouTube spam	SMS spam	Twitter spam	YouTube spam	YouTube spam
Dataset	UCI Machine Learning	Not stated	Not stated	Not stated	Not stated

Table 4 shows the comparison between previous research projects in detection of spam. There is five research projects has been listed to compare the result with accuracy of detection spam and the algorithm used. This table also shows that the most accurate in detection of spam is using Random Forest algorithm with result above 90%. Figure 2 shows how Random Forest works.

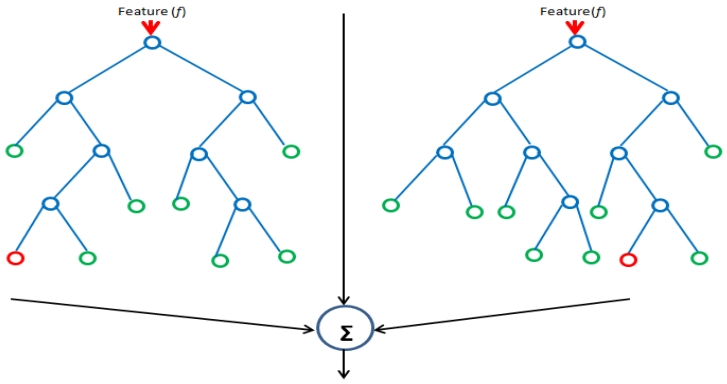

Figure 2. Random Forest model

Random Forest can give the most accurate result because it is work by built multiple of decision trees and merges it together to get stable prediction.

3. RESEARCH METHOD

In this section discuss the methodology used for video sharing spam comment feature detection. It consists of data collection, tokenization, lemmatization, feature selection and classification modules. Several experiments are conducted in order to identify the most suitable technique to detect spam comment. The performance evaluation used in this research is Accuracy.

There is two modules which is Module 1: Data Collection and Module 2: Text Mining. Figure 3 shows framework used in this work.

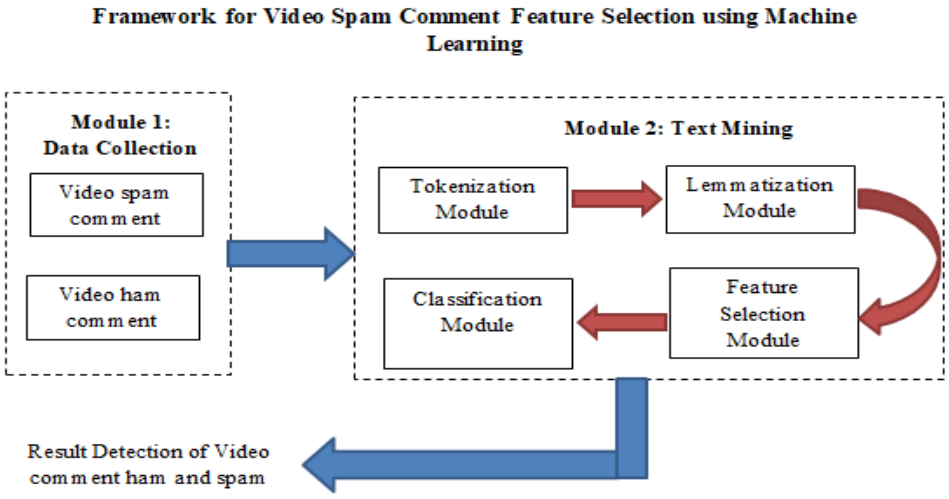

Figure 3. Framework used in Video Spam Comment Features Selection using Machine Learning Technique

3.1. Datasets

Data collection for this work is used for conducting experiments. In order to detect spam comment, a collection of spam and ham comment must be selected from the UCI Machine Learning repository. Because of time constraint to collect primary data, the existing spam dataset that has been collected by the previous researchers were chosen for this work [1]. There are 350 real comments extracted from a video and it is divided into two which is 175 comments were spam and another 175 comments were ham comments.

3.2. Tokenization

The purpose of tokenization is to split the video comment into individual words in order to smoothen out the lemmatization process. For this work, tokenization has been done using Microsoft Excel.

3.3. Lemmatization

Lemmatization is a process of grouping the similar words. For this work, the process is instead used to group words that exactly same. It is because in most cases, the video comment attacker will simply use different abbreviations of words. This process is done by using Rapid Miner.

3.4. Feature Selection

Feature Selection is important for spam comment detection. It is because the accuracy of detection spam comments depends on the features that has been selected [13]. In this experiment, only datasets that contain texts of comments is used. The features that have been extracted and evaluated for this works are bag-of-word model. The features are selected based on comparison that have been stated in related work.

3.5. Classification

After extracting features, classification is tested using WEKA tool. There is six machine learning algorithm are used in this experiment which are Random Tree, Random Forest, Naïve Bayes, KStar, Decision Table and Decision Stump. Table 5 shows the classification algorithms used in this experiment. The accuracy rate has been used to compare the algorithm's performance.

Table 5. Classification Algorithms used in Work

Classification Technique	
RT	Random Tree
RF	Random Forest
NB	Naïve Bayes
K*	KStar
DTs	Decision Tree
DS	Decision Stump

4. RESULTS AND ANALYSIS

In this section, it is explained the results of research. Experiments are performed to evaluate the performance of proposed spam comment detection. The first step is selected features on basic behavior of spam and ham comments and then extracts the features from dataset to get featured vector. After extraction of features, various classifications of algorithm such as Random Tree, Random Forest, Naïve Bayes, KStar, Decision Table and Decision Stump are applied to get performance accuracy. In Table 6 show the results of proposed approach on various machine learning algorithms.

Table 6. Results of Proposed Approach on Various Machine Learning Algorithms

Feature Selection (words)	Accuracy (%)					
	RT	*RF*	*NB*	*K**	*DTs*	*DS*
1 - 39	82.00	87.14	82.57	82.86	83.71	58.86
1 - 78	84.57	89.14	83.43	84.86	68.29	63.43
1 - 117	85.43	90.29	81.74	85.14	68.29	63.44
1 - 156	86.2	90.00	83.71	85.14	76.86	65.71
1 - 195	86.86	90.57	84.00	84.58	76.86	65.71

Based on Table 6, the highest accuracy, using Random Tree classification is 86.86% by using 195 words while the low accuracy is 82% by using 39 words.

For Random Forest classification algorithm, the highest result of accuracy is 90.57%. It used 195 words and the lowest accuracy is 87.14%. It used 39 words.

The highest accuracy for Naïve Bayes is 84% with 195 words and the lowest result is 81.74% by using 117 words.

By using KStar classification, the highest accuracy is 85.14% by using 117 and 156 words respectively. The lowest accuracy, using KStar classification is 82.86% with 39 words.

For Decision Tree classification algorithm, the highest accuracy is 83.71% by using 39 words and the lowest accuracy is 68.29% by using 78 and 117 words.

Lastly, for Decision Stump classification, the highest accuracy is 65.71% by using 156 and 195 words. The lowest accuracy is 58.86% and it used 39 words to be analyzed.

5. CONCLUSION

After comparing the performance for various machine learning algorithms, Random Forest Classification gives the highest result of accuracy which is 90.57% for 1 to 195 words of features selection. The lowest accuracy is from Decision Stump Classification which is 58.86% for 1 to 39 words of features selection. So that, Random Forest Classification were achieved the best classification results with high accuracy.

This work proposed a technique for video sharing spam comments detection to overcome the problems that have been faced by user with media social. There is 195 words of features selection has been used in six machine learning algorithms which are Random Tree, Random Forest, Naïve Bayes, KStar, Decision Table and Decision Stump to get the highest accuracy of spam detection. Out of all classification algorithms, Random Forest Classification gives the best result with 90.57% accuracy.

ACKNOWLEDGEMENTS

We would like to say thank you to Universiti Tun Hussein Onn Malaysia (UTHM) and Office for Research, Innovation, Commercialization and Consultancy Management (ORICC), UTHM for kindly proving us with the internal funding GPPS Vot No H061 and Tier 1 Vot No H237.

REFERENCES

[1] T. C. Alberto, J. V Lochter, and T. A. Almeida, "*TubeSpam: Comment Spam Filtering on YouTube*," in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 138-143.

[2] R. Kaushal, S. Saha, P. Bajaj, and P. Kumaraguru, "*KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube*," in 2016 14th Annual Conference on Privacy, Security and Trust (PST), 2016, pp. 157-164.

[3] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: A Network-Based Spam Detection Framework for Reviews in Online Social Media," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 7, pp. 1585-1595, 2017.

[4] M. Alsaleh, A. Alarifi, F. Al-Quayed, and A. Al-Salman, "*Combating Comment Spam with Machine Learning Approaches*," in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 295-300.
[5] N. Azman, M. Ariff, A. Abdullah, and M. F. Nasrudin, "Experimental Approach Based on Ensemble and Frequent Itemset Mining for Image Spam Filtering", in *Journal of Telecommunication, Electronic and Computer Engineering*," vol. 10, no. 1, pp. 121-126.
[6] R. Chowdury, M. N. M. Adnan, G. A. N. Mahmud, and R. M. Rahman, "*A data mining based spam detection system for YouTube*," in Eighth International Conference on Digital Information Management (ICDIM 2013), 2013, pp. 373-378.
[7] C. Romero, M. G. Valdez, and A. Alanis, "*A comparative study of machine learning techniques in blog comments spam filtering*," in The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1-7.
[8] J. Abernethy, O. Chapelle, and C. Castillo, "Graph regularization methods for Web spam detection," *Mach. Learn.*, vol. 81, no. 2, pp. 207-225, 2010.
[9] G. Lin, N. Sun, S. Nepal, J. Zhang, Y. Xiang, and H. Hassan, "Statistical Twitter Spam Detection Demystified: Performance, Stability and Scalability," *IEEE Access*, vol. 5, pp. 11142-11154, 2017.
[10] M. Zhiwei, M. M. Singh, and Z. F. Zaaba, "*Email spam detection: A method of metaclassifiers stacking*," Proc. 6th Int. Conf. Comput. Informatics, vol. 200, no. 200, pp. 750-757, 2017.
[11] S. S. Ali and J. Maqsood, ."*Net library for SMS spam detection using machine learning: A cross platform solution*," in 2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2018, pp. 470-476.
[12] I. Dagher and R. Antoun, "*Ham-Spam Filtering Using Different PCA Scenarios*," in 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016, pp. 542-545.
[13] N. Choudhary and A. K. Jain, "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique BT - Advanced Informatics for Computing Research," 2017, pp. 18-30.
[14] Y. Yusof and O. H. Sadoon, "*Detecting Video Spammers In Youtube Social Media*," in Proceedings of the 6 th International Conference on Computing and Informatics, ICOCI 2017, no. 082, pp. 228-234, 2017.

BIOGRAPHIES OF AUTHORS

In 2017, Nabilah Alias had received a Degree in Computer Science (Information Security) from Universiti Tun Hussein Onn Malaysia. Currently, pursuing her Master Degree in Information Technology majoring in Spam Detection at Universiti Tun Hussein Onn Malaysia (UTHM).

Cik Feresa received a degree in Information Technology and Multimedia in 2006 and master in Computer Science (Information Security) in 2009 at Universiti Teknologi Malaysia (UTM). In 2017, she obtained her PhD in the field of Information Security at Universiti Teknikal Malaysia Melaka (UTeM). She started her career as a lecturer at the Department of Information Security, UTHM from November 2011. She is an active researcher and has been written and presented a number of papers in conferences and journals.

Sofia Najwa received a degree in Engineering (Bio-Medical) in 2009 and master in Engineering (Electrical-Electronic & Telecommunication) in 2011 at Universiti Teknologi Malaysia (UTM). In 2016, she obtained her PhD in the field of Information Security at Universiti Teknikal Malaysia Melaka (UTeM). She started her career as a lecturer at the Department of Information Security, UTHM. She is an active researcher and has been written and presented a number of papers in conferences and journals.