

Spam detection by using machine learning based binary classifier

Mohd Fadzil Abdul Kadir, Ahmad Faisal Amri Abidin, Mohamad Afendee Mohamed,
Nazirah Abdul Hamid

Department of Computer Science, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia

Article Info

Article history:

Received Mar 11, 2021

Revised Jan 11, 2022

Accepted Feb 5, 2022

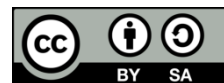
Keywords:

Binary classifier
Machine learning
Microsoft Azure
Spam detection
Text classification

ABSTRACT

Because of its ease of use and speed compared to other communication applications, email is the most commonly used communication application worldwide. However, a major drawback is its inability to detect whether mail content is either spam or ham. There is currently an increasing number of cases of stealing personal information or phishing activities via email. This project will discuss how machine learning can help in spam detection. Machine learning is an artificial intelligence application that provides the ability to automatically learn and improve data without being explicitly programmed. A binary classifier will be used to classify the text into two different categories: spam and ham. This research shows the machine learning algorithm in the Azure-based platform predicts the score more accurately compared to the machine learning algorithm in visual studio, hybrid analysis and JoeSandbox cloud.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Mohd Fadzil Abdul Kadir
Department of Computer Science, Faculty of Informatics and Computing
Universiti Sultan Zainal Abidin
Besut Campus, 22200 Terengganu, Malaysia
Email: fadzil@unisza.edu.my

1. INTRODUCTION

Today, spam has become a big problem on the internet [1]-[2]. In 2017, it was shown that spam accounted for 55% of all e-mail messages, the same as during the previous year. Spam, which is also known as unsolicited bulk email, has led to the increasing use of email for this activity as email provides the perfect way to send unwanted advertisements or junk newsgroup postings at no cost to the sender [3]-[4]. This opportunity has been extensively exploited by irresponsible organisations and resulted in the cluttering of mailboxes of millions of people all around the world. Evolving from a minor to a major concern, given the highly offensive content of the messages, spam is a huge waste of time. It also consumes a lot of storage space and communication bandwidth. End users are also at risk of deleting legitimate mail by mistake [5].

Machine learning is one field of study where computers can learn to do something without the need to be explicitly programmed for the task. The algorithm operates by building a model from input data and producing a program to perform a task such as classification [6]-[8] and prediction [9]-[12]. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages create the training dataset which will be used to fit the learning algorithm to the model in a machine learning studio. Text classification is used to determine the path of incoming mail/messages either into the inbox or straight to the spam folder. It is the process of assigning categories to text according to content. It is used to organize, structure and categorise text. It can be done either manually or automatically. Machine learning automatically classifies the text in a much faster way than

manual techniques. Machine learning uses pre-labelled text to learn the different associations between pieces of text and their output. It uses feature extraction to transform each text into a numerical representation in the form of vectors which represent the frequency of words in a predefined dictionary.

Text classification [13]-[14] is important to structure the unstructured and messy nature of such texts as documents and spam messages in a cost-effective way. Machine learning [15]-[20] can allow more accurate classification in real-time and help to improve and speed up the manual process of analysing vast amounts of data. It is especially important for a company to be able to analyse text data, help inform business decisions and even automate business processes. For example, text classification is used in classifying short texts such as tweets or headlines. It can also be used in larger documents such as media articles, social media monitoring, brand monitoring and so on.

Competition between filtering methods and spammers is being waged daily, as spammers begin to use increasingly tricky methods to overcome spam filters, such as using random sender addresses or appending random characters at the beginning or end of email subject lines. There is a lack of machine learning used in the development of models to predict this activity. Spam is a waste of time to the user since they have to sort the unwanted junk mail, and it consumes storage space and communication bandwidth. Rules in other existing sorting methods must be constantly updated and maintained, making it a burden to users, and it is hard to manually compare the accuracy of classified data.

Much research has been conducted in detecting and filtering spam email using a variety of techniques shown in Table 1. Guzella and Caminhas [21] conducted “A review of machine learning approaches to spam filtering.” In their paper, they found that Bayesian Filters that are used to filter spam require a long training period before they can function efficiently. Ananthi and Sathyabana [22] conducted research on “Spam filtering using KNN.” In this paper, she used k-nearest neighbors (KNN), as it is one of the simplest algorithms. An object is classified by a majority vote of its neighbours where the class is typically small. Sharma *et al.* [23] conducted a survey on spam detection techniques. In this paper, they found that an artificial neural network (ANN) must be trained first to categorize emails into spam or non-spam categories, starting from the particular data sets.

Table 1. Spam filtering techniques

TITLE	REMARK
SVM-based spam filter with active and online learning	Select the most useful example for labelling and add the labelled example to training set to retrain model.
Learning to classify text using support vector machines: methods, theory, and algorithms	Gaps will highly affect its completeness as a handbook in courses on machine learning for text classification and NLP
Machine learning in automated text categorisation	Text categorisation is one of the excellent methods used for checking whether a given learning technique can scale up to large size of data
Text classification using string kernels	Might be slower chunking for large data set and the quality of approximation and error generalisation is related.
Support vector machine active learning with applications to text classification	The modified existing data sets will only differ by one instance from the original labelled data set. Learning an SVM can be foreseen on the original data set.
Spam detection using text clustering	It works equivalent to SVM in form of precision on unsupervised clustering method for detecting spam.

Tong and Koller [24] conducted research on “Support vector machine active learning with applications to text classification.” In this paper, they presented a new algorithm for active learning with Support vector machines (SVM) induction and transduction. This is used to reduce version space as much as it can at every query. They found that the existing dataset only differed by one instance from the original labelled data set. Sasaki and Shinnou [25] conducted research on “Spam detection using text clustering.” They used text clustering based on a vector space model to construct a new spam detection technique. This new spam detection model can find spam more efficiently even with various kinds of mail. Mahinovs and Ashutosh [26] conducted research on “Text classification method review.” They tested the process of text classification using different classifiers, which are natural language processing, statistical classification, functional classification and neural classification. They found that all the classifiers worked well but need improvement, especially to the feature preparation and classification engine itself in order to optimize the classification performance.

In this study, the machine learning technique Vowpal Wabbit algorithm has been chosen to overcome the disadvantages of other methods. For example, real-time blackhole list techniques can generate false-positive results while Bayesian filters require a training period before they start working well. A Vowpal Wabbit algorithm is leveraged for classification of objects of different classes in order to learn the

classification rules from the messages [5]. The algorithm is provided with input and output data and has a self-learning program to solve the given task. Searching for the best algorithm and model can be time-consuming. The two-class classifier is the best way to classify a type of message as either spam or ham. This algorithm is used to predict the probability and classification of data outcome.

The paper is organized as follows. Section 2 introduces readers to the proposed method Vowpal Wabbit algorithm, and how it can be used in text classification specifically for spam email detection. Then section 3 presents the results and analysis of the implementation and lastly this work is concluded section 4.

2. PROPOSED METHOD

2.1. Vowpal Wabbit algorithm

Vowpal Wabbit (VW) is one of the machine learning systems using varied techniques, for instance, online hashing and interactive learning. VW supports multiple supervised and semi-supervised learning problems: classification, regression and active learning. It also supports multiple loss functions and regularisation. In this project, a serialiser is implemented in the deployment phase. The serialiser will traverse all properties of the VW features. It provides type based custom features as shown in Figure 1.

```

public class Classification Result
{
    public string Message {get; set;}
    public string Classification {get; set;}
    public Time Span Elapsed Time {get; set;}//period of time

    public Classification Result (string message, string classification, Time Span time)
    {
        Message = message;
        Classification = classification;
        Elapsed Time = time;
    }
}

```

Figure 1. The class row is an example of a user defined type usable by serializer

2.2. Text classification

Text classification is a supervised machine learning method which is used to classify data into one or more defined categories. It is one of the most important methods used in spam detection, filtering, categorisation and analysing real-world issues such as news articles. In this project, text classification is used to classify raw data into training and testing datasets that will be used to test the spam detection web service. Furthermore, it is used to classify the data into two classification types: spam and ham. Text classification with machine learning uses prelabelled data to make classifications that can learn the different associations between the text of particular input and expected output. Figure 2 shows the flow of text classification in machine learning.

The command line in Figure 3 shows the `RunworkerAsync()` which is the method used to start the execution of a background operation or asynchronous operation. It submits a request to start running the operation asynchronously. Hence, if an operation is already running, it will raise an exception.

In order to obtain the percentage of spam detected, all instances are calculated. The frequencies of messages, correct words, wrong words, total count, and elapsed time are considered to obtain the most accurate spam detection. The calculation in Figure 4 shows all the instances that will be calculated and shown as results at the end of testing and deploying the spam detection machine learning (ML) web service.

2.3. Spam detection ML

In theory, spam detection can be implemented at any location, and multiple stages of the process can occur at the same time. In this section, the flow of setting up spam detection ML web service is discussed, from the raw data into training it in the ML web service itself.

The raw existing data from the machine learning data repository is processed into a .csv file until ready to use. Then, the formatted data will be pre-processed using data pre-processing modules such as by removing numbers and special characters. The prepared data will then be processed using a two-class regression algorithm which is best to classify data. After the model is ready, it will be tested to score and classify data before evaluating and deploying the web service. An API key will then be generated to be used to expose the spam detection ML web service.

Classifications of the type of spam and ham are calculated based on the threshold that falls either above or below the cutover. The threshold of this data is the frequency of every single word count based on a dictionary of bigrams. For example, the cutover gained to divide the score is roughly 10.888. This cutover is calculated based on the formula in Figure 4. Hence, the classification type will be determined using the cutover. Once the threshold falls above the cutover, it will display the result as ham and, if below, then it is classified as spam. To provide a further understanding, the data flow Figure 5 shows the right flow of the spam detection machine learning process and setup.

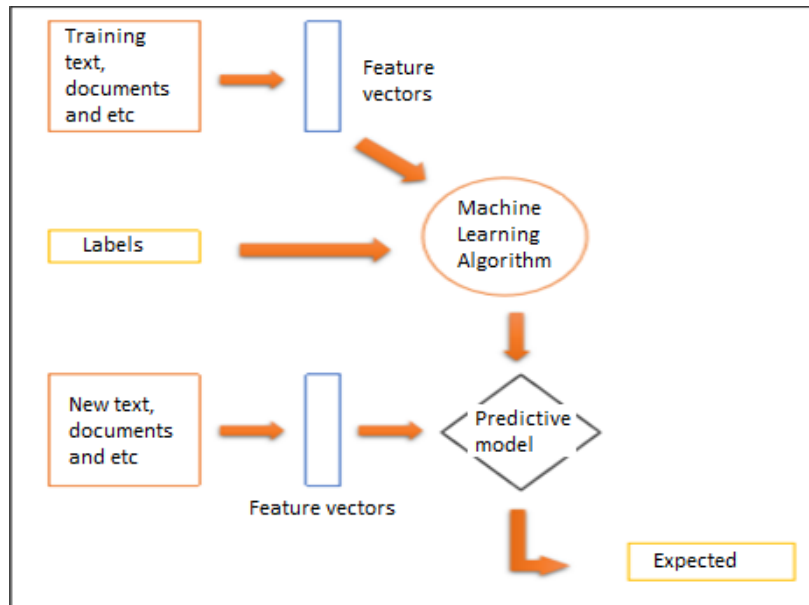


Figure 2. Machine learning-based system

```

public void ScoreClassifierAsync(string file){
    if (!worker.IsBusy)
    {
        labeledMessages = GetLabeledMessages(file);
        worker.RunWorkerAsync();
    }
}
  
```

Figure 3. Example of text classifier code

```

LabeledMessages = messages;
Correct=messages.Count(x=>x.ModelClassification
==x.RealClassification);
Total = messages.Count;
Wrong = Total - Correct;
Accuracy = (Decimal) Correct / Total;
ElapsedTime = elapsedTime;
  
```

Figure 4. A calculation used to count the score model

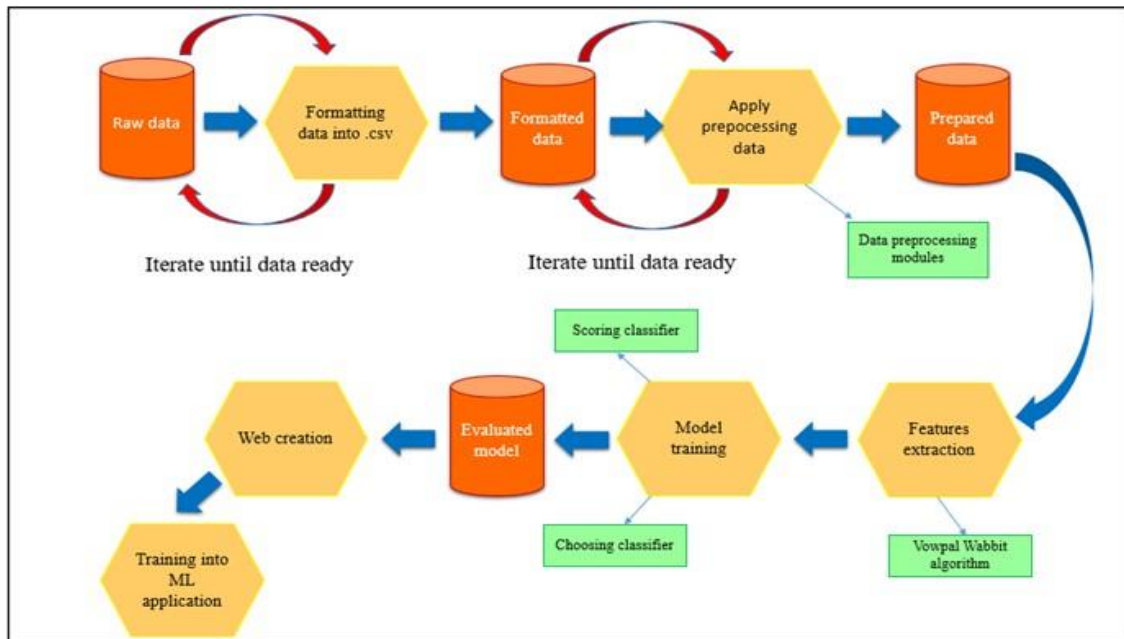


Figure 5. Data flow of spam detection machine learning

3. RESULTS AND DISCUSSION

This section presents results based on experiments conducted in this project for spam probability, elapsed time, and spam detection comparisons. By using different malware detection. All of the graphs above are focused on comparing detections using different malware detection tools.

3.1. Classification and probability

The probability of classification is measured by counting the number of spam flags. According to Peters [27], Google has examined a great number of potential factors that can predict whether a site might be penalised or banned due to spam. Each flag has its own warning sign that indicates the message as spam. Therefore, to calculate this probability, a spam score will record the number of flags triggered by the data. Hence, Figure 6 shows the relationship between the number of flags and the probability of classification type. The overall likelihood of spam increases as the number of flags increases.

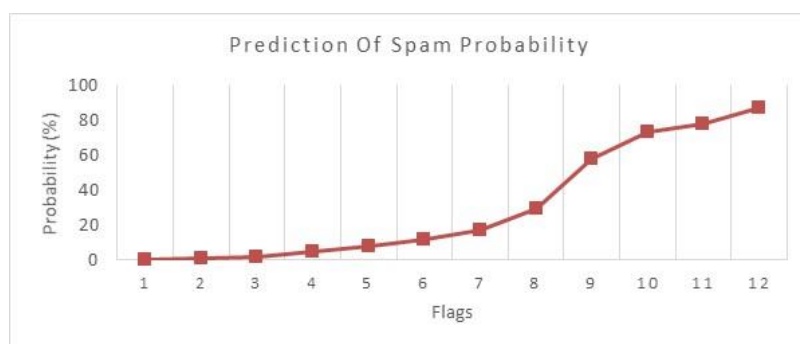


Figure 6. The relationship between number of flags and classification type

3.2. Elapsed time and message count

Elapsed time is the amount of time between the beginning and the end of the execution process. In simplest terms, elapsed time is the processing time of a process or event. In this project, both elapsed time and message count are taken into consideration in order to score the accuracy. This is to ensure the efficiency of the model by reducing the processing time even when the message count is large. Figure 7 provides a comparison of elapsed time using the same messages in four different tools.



Figure 7. The relationship between message count and time elapsed

3.3. Accuracy and message count

The message count or frequency of words is calculated in order to obtain the highest percentage of accuracy. This is because the messages are the important element to test spam detection. Figure 8 shows that all of the tools used verified that the accuracy of detection is affected by the message count.

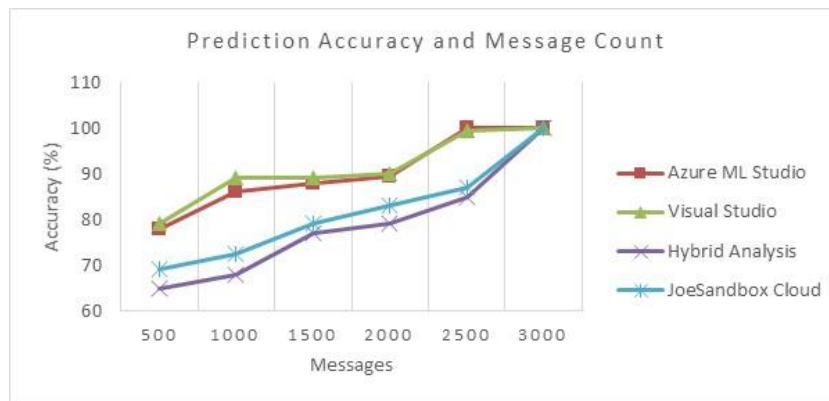


Figure 8. The relationship between percentage of accuracy and message count

3.4. Accuracy and elapsed time

The relationship between elapsed time and accuracy are also taken into consideration. Sometimes, a shorter time does not mean greater accuracy. The time affects the accuracy by processing as much data as possible and is shown in Figure 9.

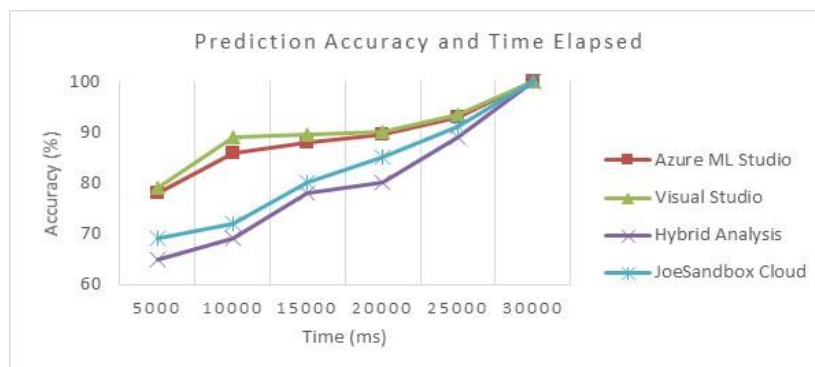


Figure 9. The relationship between accuracy percentage and time elapsed

4. CONCLUSION

The performance of a classification technique is affected by the quality of the data source. Irrelevant and redundant features of data not only increase the elapsed time, but may also reduce the accuracy of detection. Each algorithm has its own advantages and disadvantages, as stated before, supervised ML is able to efficiently separate messages and classify the correct categories. It is also able to score the models and weight them successfully. For instance, Gmail's interface uses an algorithm based on a machine learning program to keep their users' inboxes free of spam messages. During the implementation, only text (messages) can be classified and scored, not domain names and email addresses. This project focusses only on filtering, analysing and classifying messages, not blocking them. Hence, the proposed methodology may be adopted to overcome the flaws of existing spam detection. From this project, it can be concluded that Microsoft Azure machine learning studio is a cloud collaborative tool which capable to predict analytic solutions in particular data. This research has been leveraged on the Azure machine learning by modifying the Vowpal Wabbit algorithm in order to detect spam. The classification model and score weights based on words used will determine the spam.

ACKNOWLEDGEMENTS

This work was supported by the Center for Research Excellence and Incubation Management, Universiti Sultan Zainal Abidin.




REFERENCES

- [1] Y. Vernanda, S. Hansun, and M. B. Kristanda, "Indonesian language email spam detection using n-gram and naïve bayes algorithm," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 9, no. 5, pp 2012–2019, 2020, doi: 10.11591/eei.v9i5.2444.
- [2] Y. K. Zamil, S. A. Ali, and M. A. Naser, "Spam image email filtering using K-NN and SVM," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 1, pp. 245–254, 2019, doi: 10.11591/ijece.v9i1.pp245-254.
- [3] N. M. Al-Obaidi and M. M. Al-Jarrah, "Statistical keystroke dynamics system on mobile devices for experimental data collection and user authentication," in *9th International Conference on Developments in eSystems Engineering (DeSE)*, 2016, pp. 123–129, doi: 10.1109/DeSE.2016.21.
- [4] N. A. Hamid, S. Safei, S. D. M. Satar, S. Chuprat, and R. Ahmad, "Mouse movement behavioral biometric systems," in *Proceedings-2011 International Conference on User Science and Engineering, i-USEr 2011*, 2011, pp. 206–211, doi: 10.1109/iUSEr.2011.6150566.
- [5] P. U. Anitha, C. Rao, and T. Sireesha, "A survey on: e-mail spam messages and bayesian approach for spam filtering," *International Journal of Advanced Engineering and Global Technology*, vol. 1, pp. 124-136, 2013.
- [6] P. Tanwar and P. Rai, "A proposed system for opinion mining using machine learning, nlp and classifiers," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 4, pp. 726–733, 2020, doi: 10.11591/ijai.v9.i4.pp726-733.
- [7] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "Classification of instagram fake users using supervised machine learning algorithms," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 3, pp. 2763–2772, 2020, doi: 10.11591/ijece.v10i3.pp2763-2772.
- [8] F. Khan, J. Ahamed, S. Kadry, and L. K. Ramasamy, "Detecting malicious URLs using binary classification through ada boost algorithm," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, pp. 997–1005, 2020, doi: 10.11591/ijece.v10i1.pp997-1005.
- [9] J. Attenberg, K. Weinberger, A. Dasgupta, A. Smola, and M. Zinkevich, "Collaborative email-spam filtering with the hashing trick," in *Proceedings of the Sixth Conference on Email and Anti-Spam*, 2009.
- [10] W. A. Awad and S. M. Elseuofi, "Machine learning methods for spam e-mail classification," *International Journal of Computer Science & Information Technology*, vol. 3, no. 1, pp. 173-184, 2011, doi: 10.5121/ijcsit.2011.3112.
- [11] J. Barnes. "Azure machine learning." *Microsoft Azure Essentials*. 1st ed., Microsoft, 2015.
- [12] S. Juma, Z. Muda, M. A. Mohamed, and W. Yassin, "Machine learning techniques for intrusion detection system: A review," *Journal of Theoretical and Applied Information Technology*, vol. 72, no. 3, pp. 422-429, 2015.
- [13] T. H. Nguyen and K. Shirai, "Text classification of technical papers based on text segmentation," *NLDB 2013. Natural Language Processing and Information Systems*, vol 7934. Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-38824-8-25.
- [14] P. Y. Pratiksha and S. H. Gawande, "A comparative study on different types of approaches to text categorization," *International Journal of Machine Learning and Computing*, vol. 2, no. 4, pp. 423-426, 2012.
- [15] Q. Wang, W. L. Li, and Z. Z. Jin, "Review of text classification in deep learning," *Open Access Library Journal*, vol. 8, p. e7175, 2021, doi: 10.4236/oalib.1107175.
- [16] J. Kolluri, S. Razia, and S. R. Nayak, "Text classification using machine learning and deep learning models," *International Conference on Artificial Intelligence in Manufacturing & Renewable Energy (ICAIMRE) 2019*, 2019 doi: 10.2139/ssrn.3618895.
- [17] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," In: Aggarwal C., Zhai C. (eds) *Mining Text Data*, Springer, Boston, MA, 2012, doi: 10.1007/978-1-4614-3223-4_6.
- [18] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: a survey," *Information* 2019, doi: 10.3390/info10040150.
- [19] R. Rosly, M. Makhtar, M. K. Awang, M. I. Awang, and M. N. A. Rahman, "Analyzing performance of classifiers for medical datasets," *International Journal of Engineering and Technology (UAE)*, vol. 7, no. 2, pp. 136-138, 2018, doi: 10.14419/ijet.v7i2.15.11370.
- [20] M. Iqtait and F. S. Mohamad, "Multiple classifiers for age prediction against AAM and ASM," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 456-461, 2019, doi: 10.35940/ijrte.B1080.0782S319.




- [21] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10206-10222, 2009, doi: 10.1016/j.eswa.2009.02.037.
- [22] S. Ananthi and S. Sathyabama, "Spam filtering using KNN," *Journal of Computer Applications*, vol. 2, no. 3, pp. 20-23, 2009, doi: 10.5120/ijca2016908471.
- [23] A. Sharma, Manisha, D. Manisha, and D. R. Jain, "A survey on spam detection technique," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 12, pp. 8688-8691, 2014, doi: 10.17148/IJARCCCE.
- [24] S. Tong, and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, vol. 2, pp. 45-66, 2001, doi: 10.1162/153244302760185243.
- [25] M. Sasaki and H. Shinnou, "Spam detection using text clustering," *International Conference on Cyberworlds (CW'05)*, 2005, p. 319, doi: 10.1109/CW.2005.83.
- [26] A. Mahinovs and A. Tiwari, "Text classification method review," *Decision Engineering Report Series* (Ed. R Roy and D Bexter). Cranfield University, 2007.
- [27] M. Peters. *A new analysis of google SERPs across search volume and site type*. 2013. Accessed: December 3, 2021. [Online]. Available: <https://moz.com/blog/google-serps-across-search-volume-and-site-type>

BIOGRAPHIES OF AUTHORS






Mohd Fadzil Abdul Kadir    received Ph.D. in Engineering (System Engineering) from the Mie University, Mie, Japan, in 2012. Since 2006, he has been with the Faculty of Informatics & Computing, Universiti Sultan Zainal Abidin, where he is currently a Senior Lecturer. His main areas of research interest are Digital Image Processing, Pattern Recognition, Information Security and Cryptography. He is also a member of the Malaysia Board of Technologists. He can be contacted at email: fadzil@unisza.edu.my.






Ahmad Faisal Amri Abidin    is a Senior Lecturer of Computer Science at Universiti Sultan Zainal Abidin. His research interests are secure protocols, trust computing and trust in social networks. He has worked extensively on the trust in generic social networks, which he leveraged and integrated the sixth degree of separation theory to gain trust. He maintains interests in various aspects of hacking and security countermeasures including confidentiality, integrity and availability of data and information systems. Currently, he directs Computer Security and Networking Lab and is affiliated with the Facilities and Technical Management of Faculty of Informatics and Computing. He can be contacted at email: faisalamri@unisza.edu.my.



Mohamad Afendee Mohamed    received his PhD in Mathematical Cryptography in 2011 and currently serves as an associate professor at Universiti Sultan Zainal Abidin. His research interests include both theoretical and application issues in the domain of data security, and mobile and wireless networking. He can be contacted at email: mafendee@unisza.edu.my.



Nazirah Abd Hamid    is a lecturer in University Sultan Zainal Abidin, Terengganu. She holds a degree in Bachelor of Information Technology from University Utara Malaysia (UUM), in 2004 and M. Sc. Com. (Information Security) from University Teknologi Malaysia (UTM) in 2011. Her research interests are Information Security and Cyber Security, Pattern Recognition and Data Mining and Artificial Intelligence. She can be contacted at email: nazirah@unisza.edu.my.