2015 IEEE 2015 International Conference on Computer, Communication, and Control Technology (I4CT 2015), April 21 - 23 in Imperial Kuching Hotel, Kuching, Sarawak, Malaysia

An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification

Masurah Mohamad
Software Engineering Research Group (SERG)
Universiti Teknologi Malaysia
81310 UTM, Johor Bahru, Johor, Malaysia
masue_10@yahoo.com

Ali Selamat
Software Engineering Research Group (SERG)
Universiti Teknologi Malaysia
81310 UTM, Johor Bahru, Johor, Malaysia
aselamat@utm.my

Abstract—**In this paper, a spam filtering technique, which implement a combination of two types of feature selection methods in its classification task will be discussed. Spam, which is also known as unwanted message always floods our electronic mail boxes, despite a spam filtering system provided by the email service provider. In addition, the issue of spam is always highlighted by Internet users and attracts many researchers to conduct research works on fighting the spam. A number of frameworks, algorithms, toolkits, systems and applications have been proposed, developed and applied by researchers and developers to protect us from spam. Several steps need to be considered in the classification task such as data pre-processing, feature selection, feature extraction, training and testing. One of the main processes in the classification task is called feature selection, which is used to reduce the dimensionality of word frequency without affecting the performance of the classification task. In conjunction with that, we had taken the initiative to conduct an experiment to test the efficiency of the proposed Hybrid Feature Selection, which is a combination of Term Frequency Inverse Document Frequency (TFIDF) with the rough set theory in spam email classification problem. The result shows that the proposed Hybrid Feature Selection return a good result.**

Keywords—Spam, filtering, algorithm, feature selection, TFIDF, rough set theory

I. INTRODUCTION

Spam is strongly disliked by internet users, especially for those using electronic mail. It floods and attacks the inbox by sending an unwanted message via text, image, video and also voice file formats. This unsolicited or unwanted message actually waste our time, resources and sometimes affect the emotion, as the users need to keep on removing or responding to the spam messages. Besides, spam also causes a huge crisis for companies and organizations.

According to [1], spam emails are classified into six major types, which are email spam, instant messenger spam, unsolicited text message, comment spam, junk fax and social networking spam. Noormadinah et. al [2] claimed that all types of communication including phone conversation, instant messaging, short message service (sms), video call, tele-conference and email are misused by spammers to send spam in order to gain profit at no cost at all. Different techniques have been applied to prevent spam. Moreover, anti-spam law has also been enacted by the government for spammers who continue distributing unwanted messages or advertisement for internet users [3]. Researchers and email hosting providers had identified several techniques and mechanisms to avoid spam from flooding the email inbox.

There are two major types of spam filtering system currently in used: machine learning and non-machine learning [4]. Heuristic and blacklisting are most commonly used for non-machine learning technique. Author [2] highlighted that, among all of these techniques, the most successful technique, which can really block spam uses the machine learning algorithm. The reason why the machine learning technique is most preferred by researchers is its high accuracy in blocking spam emails [2], [3].

As mentioned in [1], there are two main categories of spam filtering techniques for machine learning. Both of categories have their own benefits and weaknesses. These two techniques are:

- Content Based Spam Filtering, and
- Non-content based or Metadata Based Spam Filtering such as HTML tags.

According to these two categories, a number of machine learning algorithms proposed by researchers are listed in table 1 below.

TABLE 1: Examples of machine learning techniques [1-5].

Name of Techniques	Type of Techniques
Naïve Bayesian (keyword based)	Content based
Memory-based approach	Content based
Support Vector Machine	Content based
Lazy learning algorithm	Content based
Header information	Non-content based
Behavior based features	Non-content based
Back-propagation neural network	Non-content based

On top of that, we also take the initiative to investigate the ability of one of the machine learning algorithms, called the rough set theory with the help of hybrid feature selection as a feature selection method in classifying spam email. The entire organization of this paper is as follows: In the second section, an overview of related research for spam filtering techniques is discussed. The third section provides the experimental work to test the ability of the rough set theory in the classification task with the implementation of hybrid feature selection. In section four of this paper, the result of

978-1-4799-7953-3/15/$31.00 ©2015 IEEE

the experimental work is discussed, and lastly, in section five, a conclusion of the entire work with future research direction is highlighted.

II. SPAM FILTERING METHODS

Generally, in the classification task, data cleaning or pre-processing need to be done before being classified by the classifier [5]. This process will make sure that the features to be analyzed and taken into account can truly help in generating a good result. This section will discuss the highlighted techniques applied in the filtering spam task namely, feature selection methods and machine learning approaches.

A. Feature Selection Methods

Feature selection methods are used to overcome the task of extracting high dimensional data into the smallest-possible [2]. As mentioned in [2], [5], Information Gain (IG), Gini Index and $X^{2\text{-Statistic}}$, Fuzzy Adaptive Particle Swarm Optimization (FAPSO) and Term Frequency Inverse Document Frequency (TF-IDF) are among the popular methods used in spam filtering task.

1. Information Gain (IG)

It is used to measure the amount of information which can be provided to the classification system. Larger value of Information Gain (IG) increases its significance [2].

2. Gini Index

It is a non-purity split method, which had been improved from a decision tree induction [5]. This method considers feature containing the least category of information in every message.

3. $X^{2\text{-Statistic}}$

This method is also known as the Chi-square test, which is used to test the independence of two variables in mathematical statistics. It is applied in the feature selection method in order to determine the independence of a feature ti, and a class cj. If $X^2(t_i,c_j)=0$, feature t_i, and class c_j are independent, feature t_i does not contain any category information. Otherwise, greater value of $X^2(t_i,c_j)$ indicates more category information owned by feature t_i [5].

4. Fuzzy Adaptive Particle Swarm Optimization (FAPSO)

FAPSO is divided into three levels, which are core feature subset selection, feature subset selection and spam filtering. The objective of this proposed technique is to identify an optimal feature subset [2].

5. Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF is a technique from mathematics. It identifies the frequency of words in a document by calculating the value of relevant words through an inverse ratio of the word's frequency in a document to the percentage of documents the words appears in [6]. TF-IDF returns high values of percentage if the words are common in a single document or a small group of documents.

B. Machine Learning Approaches

Machine learning approaches are usually applied in the spam filtering task. There are able to extract the knowledge taken from the supplied original dataset into information in any classification task [7]. Besides, these approaches are also able to improve their performance through the learning experience.

Naïve Bayes is one of the machine learning algorithms, which is always applied by researchers. As mentioned in [6], this algorithm was initially proposed by Sahami, Dumais, Heckerman and Horvitz and also extended by Graham. This proposed classifier classifies the problem by implementing a decision theoretic framework in the classification process. This algorithm also outperformed a keyword-based as experimented by Androutsopoulos, Koutsias, Konstantinos and Spyropoulos as highlighted by [4]. Another popular machine learning approach frequently implemented by researchers is the Support Vector Machine (SVM). It was first applied by Drucker, Wu and Vapnik, using the Bag of Word (BoW) representation with binary, frequency or *TF-IDF* features, selected according to information gain and two private corpora [4], [6]. Another machine learning approach that has been applied is the Lazy Learning to recognize concept drift in keyword-based spam filters.

Rough set is also a common technique applied generally in data mining problem and specifically in classification. Rough set theory is a mathematical tool introduced by Pawlak in 1982 in order to deal with the vagueness and uncertainty of information [4]. It can be used as a feature selection technique [8] or even as a classifier in the classification task. The results of a rough set approach are usually presented in the form of a set of decision rules derived from a decision table. As mentioned in [9], [10], many researchers have proved that the rough set theory is very effective when applied in many data mining applications, especially when dealing with numerous attributes. The rough set theory will reduce irrelevant and redundant words from a large database.

Even though these machine learning approaches have provided a number of successful classification results, there are still few challenging problems, such as in analyzing non-content based keywords. The most applied non-content based keywords during a classification process are the email header sections and spam behaviors [4]. To make use of these two beneficial techniques, a machine learning and non-content based keywords; rough set spam detection system is proposed.

III. PROPOSED ROUGH SET SPAM DETECTION SYSTEM

The main objective of the proposed system is to enhance the current machine learning approach in detecting spam email and improving the classification accuracy [11]. This system implements a combination of two feature selection methods, TF-IDF and the rough set feature reduction. As mentioned in section 2 previously, both TF-IDF and rough set feature reduction are good as feature selection techniques in a classification process. However, these two techniques sometime will return poor result of classification because of inadequate data or information if they work individually [6]. Therefore, this proposed hybrid method hope to increase the classification task accuracy rate because these two techniques will facilitate the classifier in generating a good filtering result. The following section will discuss the processes of spam detection, starting from dataset collection until classification result. Fig. 1 depicts

the proposed framework for this experimental work. This proposed framework had three main activities, namely pre-processing task of image and text email, feature selection process and classification task. Each of these activities is explained briefly in the following section.

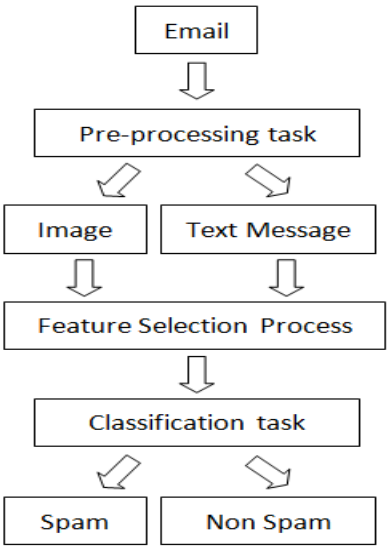


Fig 1: Proposed Spam Filtering Framework

IV. EXPERIMENTAL WORK

Experimental work was done to test the ability of the chosen machine learning approach, which is the rough set theory in spam filtering task as a classifier. This experimental work also tested the capability of a hybrid feature selection, which is the combination of Term Frequency Inverse Document Frequency (TF-IDF) and the rough set theory in helping the rough set classifier to classify spam messages.

First, the dataset comprising of text messages and images were collected from our own email inbox and several web pages. Even though there are many public spam datasets provided and used by other researchers [6], we preferred to use our own collection, where the content of the public dataset is nearly similar with our own collection and to preserve the originality of research work. Furthermore, the type of data collection is not the major concern in this research as long as the content of an email contains "spam text and spam image". This experimental work had collected 169 emails comprising of texts and images were converted into text files, where 114 text files were categorized as spam, while another 55 text files were categorized as ham. These emails were divided into two parts, whereby 60% were used as training data and 40% were used as testing data.

Secondly, all of these messages went through a pre-processing task prior to training and testing processes. The pre-processing task is also known as the feature extraction process. At this level, all messages were cleaned up in order to remove unnecessary words using the stop word removal and porter stemming algorithms. This process considered two languages, English and Malay.

However, an additional pre-processing task for image dataset was first required before it the pre-processing task. Many image characteristics had to be considered during the cleaning process such as color, shape, pixel, metadata, image format, edge and texture [12], [13]. In this experimental task, the embedded text was considered and recognized using Optical Character Recognition (OCR) technique. Other image characteristics were ignored as they did not contribute to this experimental work.

Thirdly, after all messages were cleaned up, they went through a feature selection process. In this process, a hybrid feature selection method was applied, which combined the Term Frequency Inverse Document Frequency (TF-IDF) technique and the rough set theory. The TF-IDF value was calculated for all words in each document as an input value for the rough set theory.

A toolkit, the Rough Set Exploration System (RSES) version 2.2.2, developed by Logic Group, Institute of Mathematics, Warsaw University, Poland [14] was used to remove irrelevant and redundant words from the dataset. This application was used as a feature selection processing tool and also as a classifier in the email classification task. During the feature selection process, rules had to be first constructed manually by the user, or automatically by the system. The feature selection process, reduction process and rules generation results generated by RSES are illustrated in the Fig. 2, 3 and 4. These processes were done during the training process.

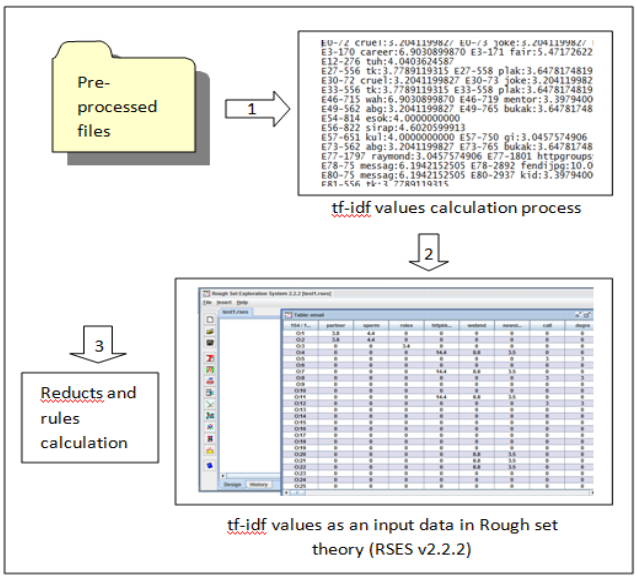


Fig 2: Feature selection process.

As depicted in Fig. 3, the attribute reduction process was done using the Genetic Algorithm method. 10 number of reduct sets (results of reducing unwanted attributes or words) were generated, where the size of the words or attribute was more than 10. Attribute reduction is a process of identifying an optimal subset of all words according to some categories. The advantages of this process are to increase classification accuracy, minimize processing time and also simplify classification results [10]. Based on these reduct sets operation, 234 rule sets were developed as shown in Fig 4. below.

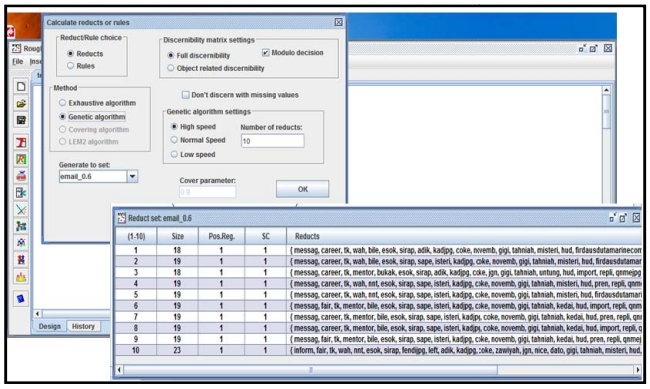


Fig 3: Attribute reduction process.

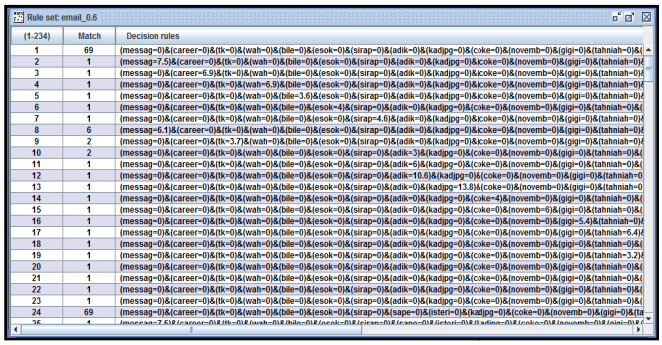
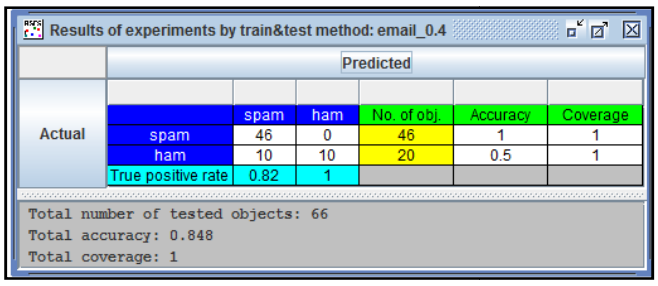


Fig 4: Rule sets generated from the process of reduct sets calculation.

Finally, the generated rules were used in the testing process by the rough set classifier to filter the emails in the classification task. 66 emails, which were a combination of 46 spam emails and 20 ham emails were used in the testing process. As predicted earlier, the rough set theory gave a good result, which achieved 0.848 or 84.8% accuracy as shown in figure 5.



Results of experiments by train&test method: email_0.4

Actual \ Predicted	spam	ham	No. of obj.	Accuracy	Coverage
spam	46	0	46	1	1
ham	10	10	20	0.5	1
True positive rate	0.82	1			

Total number of tested objects: 66
Total accuracy: 0.848
Total coverage: 1

Fig 5: Rough set classification result.

Fig. 5 shows the classification results presented by the confusion matrix, which provides ample information. The rows in the matrix correspond to the actual decision classes, which are the possible values of decision. The columns represent the decision values returned by the classifier [6]. It can be seen that two types of objects were considered, spam and ham. The constructed classifier successfully classified all 46 spam emails as spam, but not for ham emails. Unfortunately, the classifier only identified 10 out of 20 actual ham emails as ham, while another 10 were mistakenly classified as spam.

The right side of the confusion matrix represents additional information from the classification process as described and illustrated in figure 6 below:

- No. of obj. – the number of objects (emails) in the dataset for each decision class, either spam or ham. As mentioned earlier, 66 emails were used as testing data set, which was divided into 46 spam and 20 ham emails.
- Accuracy – the ratio of correctly classified objects (emails) from the class (spam or ham) to the number of all emails assigned to the class by the classifier. Fig. 6 depicts that the accuracy of Rough set classifier to filter the spam is 1 which is equals to 100% while as ham only reached 50% of accuracy rate. The classifier was wrongly classified 20 ham emails as spam emails. This result was also influenced by the number of reducts, which were 10 as shown in figure 3.
- Coverage – the ratio of classified emails by the classifier from the class to the number of all objects in the class.

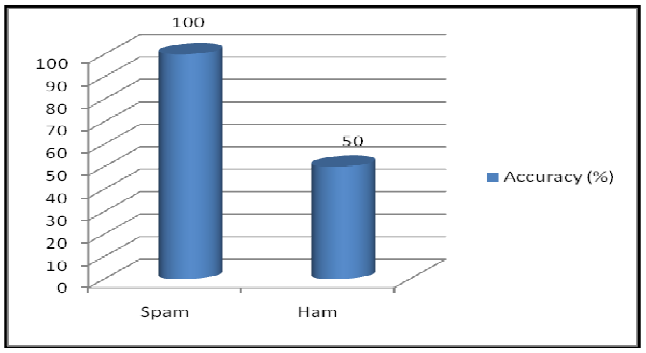
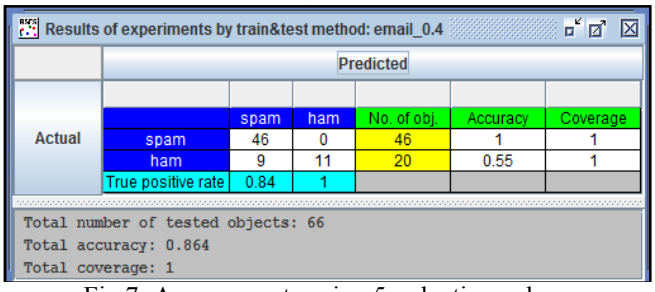


Fig 6: Classification Accuracy

The above figures Fig. 7, Fig. 8 and Fig. 9 depict the differences in the classification results when the number of reducts was changed to five, 15 and zero as comparisons. As shown in Fig. 7, 8 and 9 above, different number of reducts returned different accuracy rate. Among these four tests, it can be concluded that only five reducts was sufficient for this classification case.

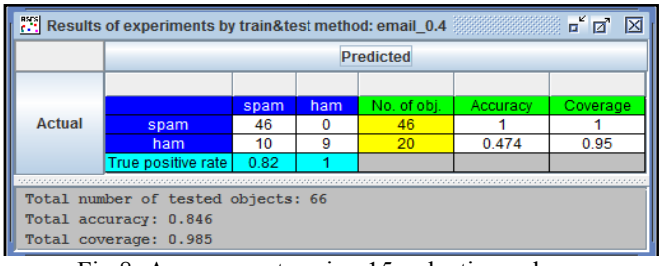


Results of experiments by train&test method: email_0.4

Actual \ Predicted	spam	ham	No. of obj.	Accuracy	Coverage
spam	46	0	46	1	1
ham	9	11	20	0.55	1
True positive rate	0.84	1			

Total number of tested objects: 66
Total accuracy: 0.864
Total coverage: 1

Fig 7: Accuracy rate using 5 reduction values



Results of experiments by train&test method: email_0.4

Actual \ Predicted	spam	ham	No. of obj.	Accuracy	Coverage
spam	46	0	46	1	1
ham	10	9	20	0.474	0.95
True positive rate	0.82	1			

Total number of tested objects: 66
Total accuracy: 0.846
Total coverage: 0.985

Fig 8: Accuracy rate using 15 reduction values

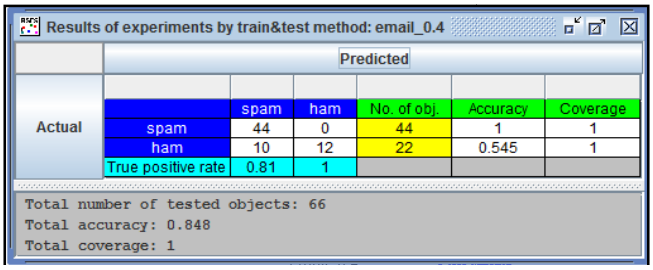


Fig 9: Accuracy rate without reduction process.

Table 2 and Fig. 10 below described the comparisons between all four test cases. Cases 1 and 3 achieved the same classification result. It shows that 10 reducts was the default value in the RSES application during the reduction process.

TABLE 2: Accuracy rate for different number of reducts.

Case #	Number of reducts	Accuracy	Coverage
1	none	84.8%	100%
2	5	86.4%	100%
3	10	84.8%	100%
4	15	84.6%	98.5%

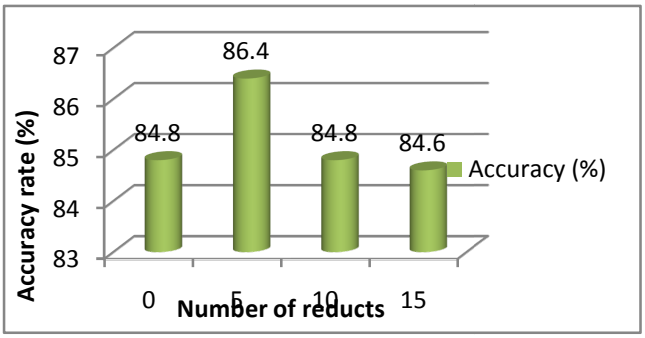


Fig 10: Classification results for the different cases.

The Fig. 11, compared the result of classification between the proposed feature selection (TF-IDF and Rough set-reducts calculation), with other two selected feature selection methods, TF-IDF with decision tree and TF-IDF and Rough set-rules generation. It shows that, combination of TF-IDF with decision tree as a feature selection method helps to generate more accurate result compared to the proposed method and TF-IDF method.

V. CONCLUSION

Feature selection is one of the important tasks in classification process. This paper has presented a hybrid feature selection method, namely the Hybrid Feature Selection, which integrates the term frequency inverse document frequency (TF-IDF) and the rough set theory to increase the classification result, generally, and email filtering, specifically. As predicted, the implementation of TF-IDF and rough set returned a reasonable good result. It shows that TF-IDF and rough sets were able to work together in order to generate concise and more accurate results. For the future work, we planned to apply another hybrid classifier such as the combination of TF-IDF with Support Vector Machine (SVM) or implementing additional methods in the pre-processing task in order to increase the classification accuracy. Besides, a comprehensive study on the overall functions of RSES is suggested for researchers who want to investigate this test case further. Finally, the use of dataset should also be considered in the classification task as different types of dataset will produce different rates of accuracy for the classification process.

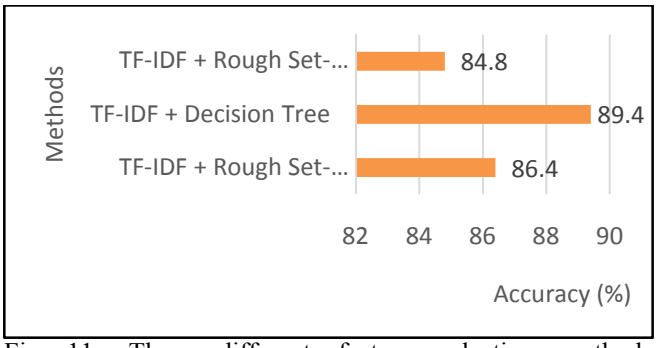


Fig 11: Three different feature selection methods classification accuracy

REFERENCES

[1] K. Jain, "A Hybrid Approach for Spam Filtering using Support Vector Machine and Artificial Immune System," pp. 5–9, 2014.

[2] N. Allias, "A Hybrid Gini PSO-SVM Feature Selection Based on Taguchi Method□: An Evaluation on Email Filtering," 2014.

[3] I. Idris, A. Selamat, and S. Omatu, "Hybrid email spam detection model with negative selection algorithm and differential evolution," *Eng. Appl. Artif. Intell.*, vol. 28, pp. 97–110, 2014.

[4] Y. Hu, C. Guo, E. W. T. Ngai, M. Liu, and S. Chen, "A scalable intelligent non-content-based spam-filtering framework," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8557–8565, 2010.

[5] J. Yang, Y. Liu, Z. Liu, X. Zhu, and X. Zhang, "A new feature selection algorithm based on binomial hypothesis testing for spam filtering," *Knowledge-Based Syst.*, vol. 24, no. 6, pp. 904–914, 2011.

[6] J. Ramos, J. Eden, and R. Edu, "Using TF-IDF to Determine Word Relevance in Document Queries," *Processing*, 2003.

[7] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to Spam filtering," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10206–10222, 2009.

[8] Y. H. -, L. D. -, D. X. -, and S. W. -, "A Novel Discrete Artificial Bee Colony Algorithm for Rough Set-based Feature Selection," *Int. J. Adv. Comput. Technol.*, vol. 4, no. April, pp. 295–305, 2012.

[9] Y. C. Hu, "Rough sets for pattern classification using pairwise-comparison-based tables," *Appl. Math. Model.*, vol. 37, no. 12–13, pp. 7330–7337, 2013.

[10] R. Li and Z. Wang, "Mining classification rules using rough sets and neural networks," *Eur. J. Oper. Res.*, vol. 157, pp. 439–448, 2004.

[11] D. Miao, Q. Duan, H. Zhang, and N. Jiao, "Rough set based hybrid algorithm for text classification," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9168–9174, 2009.

[12] B. Fadiora, "Combining Optical Character Recognition (OCR) and Edge Detection Techniques to Filter Image-Based Spam," *African J. Comput. ICT January*, vol. 5, no. 1, pp. 59–68, 2012.

[13] L. X. Mang, H. Jung, H. Y. Youn, and U. Kim, "AN INCREMENTAL LEARNING BASED FRAMEWORK," vol. 4, no. 1, 2014.

[14] http://logic.mimuw.edu.pl/~rses/