

Research Article

K-Means Clustering Scheme for Enhanced Spam Detection

^{1,2}Nadir Omer Fadh Elssied and ¹Othman Ibrahim

¹Faculty of Computing, University Technology Malaysia, 81310, Skudai, Johor Bahru, Malaysia

²Algeraf Sharq Technical College, Khartoum, Sudan

Abstract: In recent years, the problems of increasing spam mail on the internet are becoming a serious issue and difficult to detect. Furthermore, several e-mail classification methods have been proposed and their performance is achieved. Although, Naïve Bayes classifiers (NB) has been widely used in e-mail classification and is very simple and efficient, yet the problem of improving the accuracy and reducing misclassification rate still exists. Therefore, many researches are being carried out. These studies propose a hybrid scheme for e-mail classification based on Naïve Bayes and K-means clustering to obtain better accuracy and reduce the misclassification rate of spam detection. The experiment of the proposed scheme was carried out using spam base benchmark dataset to evaluate the feasibility of the proposed method. The result of this hybrid led to enhance Naïve Bayes classifiers and subsequently increase the accuracy of spam detection and reducing the misclassification rate. In addition, experimental results on spam base datasets show that the enhanced Naïve Bayes (KNavie) significantly outperforms Naïve Bayes and many other recent spam detection methods.

Keywords: K-mean clustering, machine learning, Naïve Bayes (NB), spam detection

INTRODUCTION

Nowadays, electronic mail (E-mail) has become one of the fastest and most economical ways of communication (Caruana and Li, 2012; Sadan and Schwartz, 2011). This is due to low transmission cost, fast message delivery, accessibility and enhanced efficient communication (Chhabra *et al.*, 2010). The development of the internet and the growing use of e-mail have led to the escalation of problems caused by unsolicited e-mail, regularly referred to as spam (Guzella and Caminhas, 2009). E-mail spam refers to sending irrelevant, inappropriate and unsolicited e-mail messages to numerous people (Hayati and Potdar, 2008a; Sanz *et al.*, 2008). The purpose of e-mail spam is to advertise, promote and spread backdoor or malicious programs (Hayati and Potdar, 2008b; Mohammad and Zitar, 2011). The characteristics of e-mail spam is "a user receives unsolicited e-mail that has been sent without any concern for their identity" (Vorakulpipat *et al.*, 2012). According to an e-mail estimated statistical report, over 89 billion e-mail messages were sent per day in 2012 and this figure is expected to 143 billion by 2016 (Radicati and Hoang, 2011). However, due to spam classifier, 20% of all messages that are received by users in mailboxes are spam (Manjusha and Kumar, 2010). Thus, users need to

justify the status of e-mail whether it is spam or not spam regardless of the content (Marsono *et al.*, 2009; Vorakulpipat *et al.*, 2012). There are several critical problems connected with increasing volumes of spam that is filling users' mailboxes, wasting network resources namely storage space and e-mail bandwidth, consuming users' time for deleting all spam messages and damage to computers and laptops due to viruses (Cournane and Hunt, 2004; Lai and Wu, 2007; Sahami *et al.*, 1998). Spam mails vary significantly in content and most of them are commercial adverts (Saad *et al.*, 2012). However, new variations of spam and unknown spam are very difficult to detect. The traditional way of detecting spam based on signature is no more efficient in today's systems (Ismaila and Ali, 2012). Spam detection is a program that used to detect unwanted e-mail and avoid those e-mails from entering user's inbox (Lai and Wu, 2007; Youn and McLeod, 2007). On the other hand, to reduce spam mail completely from entering user's mailbox is very difficult as the spammers changed their approaches over time to prevent detection of the spam classification (Taninpong and Ngamsuriyaroj, 2009). With development of data mining applications such as classification and clustering shown that there are need to machine learning algorithms for applied to large scale data (Attri and Kaur, 2012). Recent research show that spam detection

Corresponding Author: Nadir Omer Fadh Elssied, Faculty of Computing, University Technology Malaysia, 81310, Skudai, Johor Bahru, Malaysia

This work is licensed under a Creative Commons Attribution 4.0 International License (URL: <http://creativecommons.org/licenses/by/4.0/>).

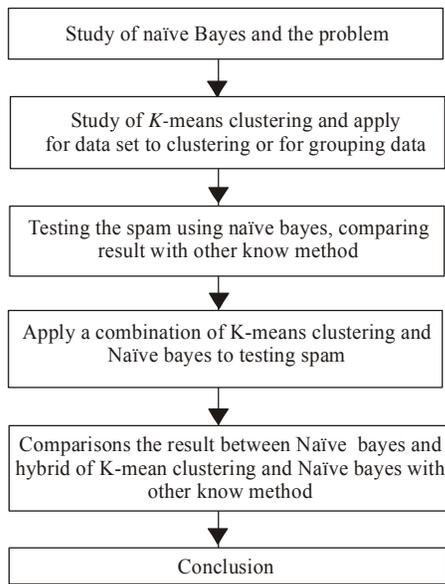


Fig. 1: Structure of the proposed method

is usually processed by Machine Learning (ML) algorithms to distinguish between non-spam and spam e-mail (Fagbola *et al.*, 2012; Guzella and Caminhas, 2009; Saad *et al.*, 2012). ML methods are able to extract the knowledge from a group of e-mails supplied and use the gained information in the categorization of newly received e-mail (Saad *et al.*, 2012). Then, various types of machine learning methods with different implementations are used namely supervised learning, unsupervised learning and semi-supervised learning. Supervised machine learning methods have been broadly studied and reported in the literature to improve the accuracy of e-mail filters (Hershkop and Stolfo, 2005). The goal of ML is to improve the performance of computer program through experience so as to make better decisions and solve problems in an intelligent way by using illustration data (Salehi and Selamat, 2011).

Currently, Naïve Bayes is one of the most popular algorithms for spam detection which is based on Bayes' theorem and attributes independence assumption (Yin and Chaoyang, 2011). Naive Bayes classifier is simple and has low time and memory requirements. Although the Naïve Bayes is widely used they need a lot of training samples to provide a degree of accuracy for classification (Ying *et al.*, 2010). Furthermore, low accuracy and prone to high misclassification percentages when using Naïve Bayes (Taninpong and Ngamsuriyaroj, 2009). The goal of this study is to propose a spam mail detection scheme by using hybrid of Naïve Bayes (NB) and the K-means clustering to improve the classification accuracy (effectiveness) and reduce the misclassification percentage. The novelty of

this study is there no prior work on e-mail classification via hybrid of Naïve Bayes and K-means clustering. However, it has been employed in related field. A comparative study has been carried out between spam detection using Naïve Bayes only and using a hybrid of Naïve Bayes and K-means clustering Structure. The proposed method is shown in Fig. 1.

Spam classification: Recently, spam has become one of critical issues for the internet society. From the industrial point of view, anti-spam suppliers are rolling out various interventions to reduce the amount of spam hazard (Tala Tafazzoli, 2009). Then, the interventions are deployed with different methodologies, hardware and software configurations and network structures (client and server). In contrast, from the research point of view, the soft computing technique, such as Artificial Immune System (AIS), evolutionary computation techniques namely Particle Swarm Optimization (PSO) and ant colony optimization, Differential Evaluation (DE) and Genetic Algorithm (GA) have been well incorporated in order to increase or enhance the accuracy of spam detection (Mohammad and Zitar, 2011; Wu *et al.*, 2011). This will increase the percentage of classification, which is one of most important areas in data mining (Hong, 2011). According to Yin and Chaoyang (2011) the spam e-mail has become a serious problem of side effects for free e-mail service (Cournane and Hunt, 2004). Then, still the spammers send spam messages to user's mailboxes although different methods have been proposed to combat it (Tala Tafazzoli, 2009). Spammers exploit the popularity and low cost of e-mail services to send unsolicited messages or spam, which fill users' accounts and waste valuable resources (Cournane and Hunt, 2004; Ma *et al.*, 2009; Salcedo-Campos *et al.*, 2012). In addition, spammers enhanced their methods by tailoring the e-mail with misspelled words, hyperlinks and patterns with strings of random characters to make the e-mail appear as a non-spam (Manjusha and Kumar, 2010).

The majority of researches focus on using spam methods to avoid the spam e-mail completely from entering the user's mailbox. Therefore, very difficult to prevent the spam mail from entering the user's mailbox due to the spammers changed their techniques over time (Taninpong and Ngamsuriyaroj, 2009). These methods are not able to reduce overhead, bandwidth, processing power, time and memory used by spam (Tala Tafazzoli, 2009). Classification is very important and popular method to solve the problem of spam (Attri and Kaur, 2012). There are several classification methods to prevent the flow of spam messages to user's mailboxes such as list-based and statistical-based methods. The statistical based methods used to increase the

Table 1: Shows the summary of the advantage and disadvantage of rule-based and content-based classification

Advantage	Advantages	Disadvantages
Rule-based	Less scanning time	- The unknown or new spam is very difficult to detect - High misclassification - The rules and the policies need to be updated by the administrator of the system all the times
Content-based	Using machine learning technique for classification	Spammers are aware of classification techniques and their functionality

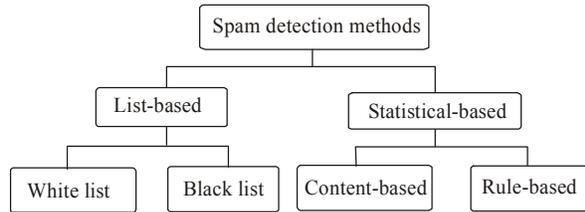


Fig. 2: Illustrates the spam detection schemes

performance of the system and it divided into two groups namely content-based (using machine learning) and rule-based (using non machine learning) (Carpinter and Hunt, 2006).

Content-based is a classification method that uses Machine Learning (ML) criteria for example naïve Bayes. Rule-based or a non-machine learning define as a set of rules to identify whether an incoming messages is spam or non-spam. Table 1 demonstrates the summary of the advantages and disadvantages of rule-based and content-based classification. In most machine learning methods, classification methods using content-based are intuitive since they are easier to implement. This technique works through some certain rules and regulations. By these rules the filter decides to pass or to block the received email. The major problem with the content and rule-based is that, the rules and the words are verified by the programmer. First, the databases and the policies need to be updated on a regular basis. Secondly, all spammers are aware of these filters and their functionality. They will try to deliver their messages using additional characters to legitimize their emails (Schryen, 2007). Finally, these techniques work after the body of the e-mail is completely received by the mail server which increases the time for checking whether the email message is spam or not (Pour *et al.*, 2012). Figure 2 illustrates the spam detection schemes. Black listing and white listing methods were implemented at the Internet Service Provider (ISP) level and sometimes white listing suffered from forgeries (Çıltık and Güngör, 2008).

LITERATURE REVIEW

Recent literature focused on spam detection to avoid spam messages (spam e-mail) from entering user’s mailboxes. The major problem in this area is how to distinguish between spam and non-spam (legitimate email). There are a number of studies which try to establish spam classification which are also known as e-

mail categorizations or spam detection system. The most current methods either prevent e-mail spam messages from the sender mail server or avoid spam from entering receiver mail server. This is because the increase of electronic commercial and electronic business leads to an increase in unwanted or unsolicited messages (e-mails). The challenge now is to develop spam classification that will effectively eliminate the increasing volume of unwanted messages automatically. Many machine learning techniques have been proposed to design the solution, some of these techniques are Naive Bayes (Youn and McLeod, 2007), Artificial Neural Network (ANN) (Vinther, 2002) and Support Vector Machine (SVM) (Xiao-Li *et al.*, 2009). Various methods used combination between two or multiples method to achieve high accuracy and low misclassification rate (error rate) rather than using individual method (Hershkop and Stolfo, 2005). In his study (Yin and Chaoyang, 2011) proposed conventional Bayesian method to reduce the influence of false negative in e-mail spam classification and also to evaluate the effective of false negative rate. The accuracy of the experimental result based on Spam Assassin standard dataset is 90%. The results lead to improve Bayesian algorithm, increase the detection rate of spam and reduce the false negative. In their study (Xiao-Li *et al.*, 2009) suggest a method of spam classification to reduce the misclassification rate and to increase the accuracy percentage using weighted SVM. He used dataset from the Chinese spam corpus ZH1 for the experiment. The experimental result is 89.50% of accuracy and shows that, the method reduces the percentage of misclassification rate while the accuracy of the classification reduced. To reduce the problem of spam emails (Youn and McLeod, 2007) presented different classifiers methods (Neural Network, SVM classifier, Naïve Bayesian Classifier and J48 classifier). The experiment was performed based on different data size and different feature size. The result is that J48 and Naïve Bayes showed the better result compared with SVM and Neural Network. The experiments with various settings are implemented in WEKA. The use of email communication is facing great threats due to unsolicited and malicious email. A combination method for spam detection to improve the accuracy and reduce the misclassification of email was proposed (Manjusha and Kumar, 2010). He used Bayesian Network to analyze the header of email while using neural network to analyze the body or content of email and using genetic algorithm as classifier trained. Zhao and Zhang

(2005) used Rough Set Classifier (RSC), in their study to reduce the misclassification rate (error rate). The result with divided the data set into 66.67 and 33.33% for training and testing respectively, for training was 97.37% of accuracy and the result of testing was 97.37%. While the result with divided the data set into 80 and 20% for training and testing respectively, for training was 92.07% of accuracy and the result of testing was 92.07%. In their study (Salehi and Selamat, 2011) proposed a hybrid between Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO) to enhance the performance of spam detection and to get better accuracy using SAIS. The result improved the accuracy of SAIS. Later (Ismaila and Ali, 2012) presented email spam classification using the spam base dataset based on combination of artificial neural network and SVM algorithm to improve the performance of detection. The testing and training result using the combinations of artificial neural network and SVM algorithm is 97.78% comparing it with 96.30 when using SVM.

In their study Taninpong and Ngamsuriyaroj, (2009) suggest an incremental adaptive spam mail classification based on multinomial Naïve Bayes classification and computed probability percentages to select the best feature to handle the problem of new spam and to reduce the misclassification percentage. The testing and training result using two different dataset namely Trec05-1 and Trec06p are 91.75 and 93.53 of accuracy, respectively.

Hitherto, no one has reached the optimal solution in order to increase the detection accuracy. There are shortcomings regarding the problem of increasing the accuracy of spam detection and decrease the misclassification. Numerous studies have shown that combining classifiers yields better results than an individual classifier. This study implements the hybrid of Naïve Bayes and K-means clustering to enhance Naïve Bayes and to increase the percentage of spam detection.

PROPOSED METHODOLOGY

In this section, we present our proposed work to meet the objective of this scheme. Different data mining classifiers are used to meet the objective of this research study. Mainly K-means clustering and Naïve Bayes (NB) classification algorithm are considered to classify the spambase dataset. The spam base dataset was partitioned into 70% of the training dataset and 30% of the testing dataset and use comma separated value (CSVed) tools for preprocessing. The dataset is applied for the two classifiers to build a mechanism. K-means clustering is used to divide dataset element (input data) into groups so that items in the same group are as similar as possible. After that the researcher applied Naïve Bayes to the classifier to improve the accuracy of detection and the steps of proposed method,

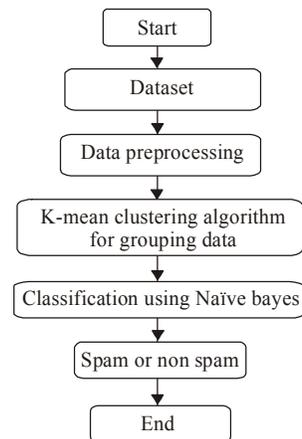


Fig. 3: The step of proposed method

see Fig. 3 (demonstrates K-means clustering and Naïve Bayes step).

Naïve Bayes algorithm: A Bayesian network is a directed acyclic graph that compactly represents a probability distribution (Sahami *et al.*, 1998). Naïve Bayes can be defined as Bayesian theorem's" from Bayesian statistic" with a conditional independency assumption that all variables or features in given category C example are conditional independent of each other (Arun and Durga, 2009; Hong, 2011). It is called Naïve Bayes because it is based on "Bayes rule" and it is relatively a new method that has quickly gained popularity because of the appropriate results that have been achieved in a wide variety of machine learning problems and because it has solid theoretical underpinnings in statistical learning theory (Long *et al.*, 2011; Salcedo-Campos *et al.*, 2012). Naïve Bayes also is a text classification algorithm and is composed of two methods namely multivariate Bernoulli and multinomial event model (Schneider, 2003) which is one of strong algorithm for categorization (Zhang and Li, 2007) and a probabilistic learning method based on statistical learning theory (Arun and Durga, 2009). There are many advantages in Naïve Bayes classification approach namely low storage space. This is because Bayesian only need to store the number of words rather than the email. Furthermore, Bayesian classification approach is suitable as a user classifier as each user can customize the classifier. Naïve Bayes is able to deal with a small number of features and takes short training and classification time. However, the drawback of Bayesian classification method is that it is weak in self-learning and self-adaptability and needs a lot of training samples. In addition, Bayesian network has the limitation of constant updating of keywords according to spammer ongoing challenges (Manjusha and Kumar, 2010). Naïve Bayes has high accuracy for binary classification, but the level of misclassification

of non-spam message is very high-that means low precision (Xiao-Li *et al.*, 2009).

Validation of each incoming email is attained by tokenizing the email and using the recalculated spam probability of each feature to classify the incoming email as spam or non-spam using following naïve:

Suppose the set of email is D, $D = [d_1, d_2, \dots, d_n]$

The set of word is $T \in d_i, i = 1, 2, \dots, n, T = [T_1, T_2, \dots, T_n]$

Type of Varsity is C (class label), $C = [\text{spam}, \text{non-spam}]$:

$$P(\text{Spam}/T) = \frac{P(T/\text{Spam}) \cdot P(\text{Spam})}{P(T)}$$

$$P(T) = P(\text{Spam}) \cdot P\left(\frac{T}{\text{Spam}}\right) + P(\text{nonSpam}) \cdot P(T/\text{nonSpam})$$

Thus,

$$P\left(\frac{\text{Spam}}{T}\right) = \frac{P\left(\frac{T}{\text{Spam}}\right) \cdot P(\text{Spam})}{P(\text{Spam}) \cdot P\left(\frac{T}{\text{Spam}}\right) + P(\text{nonSpam}) \cdot P(T/\text{nonSpam})}$$

$P(\text{Spam}/T)$ = The probability that the message is spam given that the recent message contains a given T.

$P(T/\text{spam})$ = the probability of T accruing in the message given that the message is spam:

$$P(T/\text{Spam}) = \frac{\text{Number of times the T is found in spam message}}{\text{Number of all spam messages in the database}}$$

$P(T/\text{non spam})$ = the probability of T accruing in the message given that the message is non-spam:

$$P\left(\frac{T}{\text{nonSpam}}\right) = \frac{\text{Number of times the T is found in non spam message}}{\text{Number of all non spam messages in the database}}$$

Thus for each T, the filter needs to store the number of times it has appeared in spam messages and the number of times it has appeared in non-spam messages. The total can be found through summing these two numbers. In this way $P(T/\text{spam})$ and $P(T/\text{non spam})$ can be recreated easily:

$$P\left(\frac{\text{Spam}}{T}\right) = \frac{P(\text{Spam}) \prod_{i=1}^n P(T_i/\text{Spam})}{P(\text{Spam}) \cdot \prod_{i=1}^n P\left(\frac{T_i}{\text{Spam}}\right) + P(1-\text{Spam}) \prod_{i=1}^n P(T_i/1-\text{Spam})}$$

$$P(\text{Spam}/T) + P(\text{nonSpam}/T) = 1$$

Thus,

$$P(\text{nonSpam}/T) = 1 - P(\text{Spam}/T)$$

K-means clustering: The concept of clustering has emerged for a long time. In database management, clustering data is the process of dividing data element (input data) into groups so that items in the same group are as similar as possible and items in different groups are as dissimilar as possible (DeBarr and Wechsler, 2009). Clustering is an unsupervised learning and one of the most useful methods in data mining for detection of natural groups in a dataset-means clustering algorithm and groups data based on their feature values into K clusters (Attri and Kaur, 2012). In classification the objects are assigned to predefined classes, whereas in clustering the classes are formed (Rao, 2003). There are general categories of cluster analysis methods such as tree clustering, block clustering, EM clusters and K-means clustering (Raskar and Thakore, 2011). Clustering methods may be divided into two categories based on the nature of the data and the purpose for which clustering is being used such as fuzzy clustering (each data element can belong to more than one cluster and is a mathematical method for classification such as expectation maximization method) and hard clustering (each data is divided into the distinct cluster where data elements belong to exactly one cluster such as K-means clustering) (Alguliev *et al.*, 2011; Zhang *et al.*, 2011). K-means clustering algorithm, is numerical and one of the hard clustering methods, this means that a data point can belong to only one cluster (Raskar and Thakore, 2011).

This study utilized the K-means clustering algorithm to group the messages or emails based on the similarity of their attributes or features into K disjoint groups using Euclidean distance, to improve the accuracy of spam detection. K is a positive number initialized early, before the algorithm start, to refer to the number of required clusters (Münz *et al.*, 2007). Basically, K-means clustering inspects the feature of each object, such that the objects within each cluster are similar to each other and distinct from objects in other clusters. K-means clustering is an iterative algorithm, it starts by defining an initial set of clusters and the clusters are repeatedly updated until no more improvement is possible (or the number of iterations exceeds a specified limit) (MacQueen, 1967). The steps of K-means clustering algorithm are seen in Fig. 4 (demonstrates K-means clustering step). Most of the classification techniques use distance measures for similarity such as Euclidian, cosine (similarity) and chi square (X^2) and in this study Euclidean distance is one of vector distance used for similarity. It was used to measure the difference between email messages (Euclidean distance is used as a measure to describe the similarity between data objects) (Torres *et al.*, 2009; Zhang *et al.*, 2011). Equation (1) shown below:

$$(X, y) = \sqrt{\sum_{i=1}^n (X_i - y_i)^2} \quad (1)$$

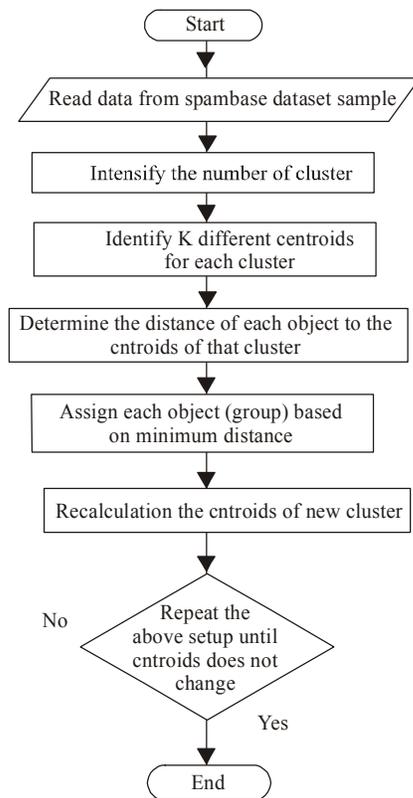


Fig. 4: k-mean clustering steps

The position of a point in a Euclidean n-space is a Euclidean vector. So, $X (X_1, X_2, \dots, X_n)$ and $Y (Y_1, Y_2, \dots, Y_n)$ are Euclidean vectors, starting from the origin of the space and their tips indicating two points.

Experimental dataset: There are various benchmark datasets available for researchers related to spam classification (Chhabra *et al.*, 2010). There has been significant effort to generate public benchmark datasets for the anti-spam classification. One of the main concerns is how to protect the privacy of the users whose non-spam messages are included in the datasets. The first approach is to use non-spam messages collected from mailing lists with public archives. There are many examples of dataset such as:

- Ling-spam
- The spam assassin
- The spam base

To develop and test spam detection, a dataset containing both non-spam and spam message is required. Furthermore, certain metrics are also required to evaluate the performance of the system. Both problems are discussed below to clarify the experimental framework considered in the study (Salcedo-Campos *et al.*, 2012). ML repository has been

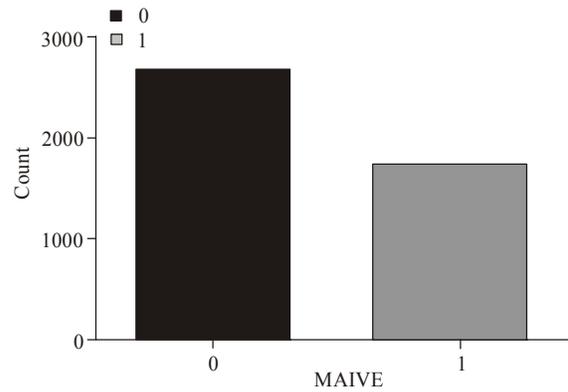


Fig. 5: Dataset analysis

used for the datasets, center for ML and intelligent system for the classifying e-mail as spam and non-spam. The spam base dataset is downloaded from the UCI machine learning repository and proposed by Hopkins *et al.* This dataset contains 58 input attributes and is composed of 4601 emails, 1813 messages (39.39%) marked as spam while the non-spam dataset has 2788 messages (60.61%) (MarkHopkins *et al.*, 1999).

This dataset was divided into two classes training and testing dataset which were divided in the ratio of 70 and 30%, respectively. Use of the hybrid of the K-means clustering and Naïve Bayes (NB) for trained classifiers by using the spam vector and the non-spam vector to detect the testing sample.

A performance index was used for the K-means clustering and Naïve Bayes (NB) to verify the effectiveness of the proposed approach. The parameter for K-means clustering and Naïve Bayes (NB) to be used in this experiment is considered as a constant change optimization process carried out by NB algorithm. Each partition use 70% as the training data and 30% as the testing data using K-means clustering and NB as the classification algorithm. This study used K-means clustering to divide the dataset into groups or clusters, then the output of clustering are used as input of classification using for NB.

Analyzing the spam base dataset from spam and the non-spam ratio as illustrates in Fig. 5, the red color indicates spam (1) while the blue color represent non-spam (0). The X-axis represented by count of email and Y-axis represented by spam type (spam or non-spam). Figure 5 shows the count of spam mails as 1813 messages while the count of non-spam mails are 2788 messages and the total emails are 4601 messages.

EXPERIMENTAL AND DISCUSSION

This section explains the accuracy result and compares this with others for training and testing results

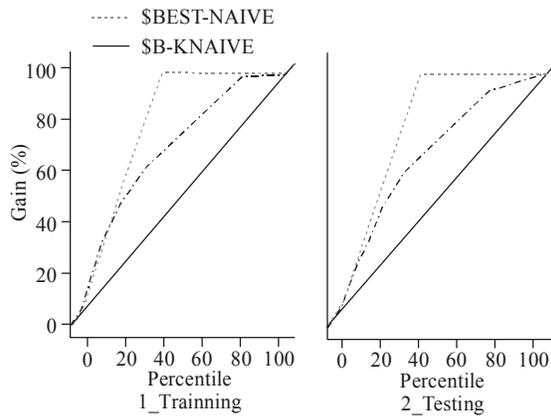


Fig. 6: Training and testing result for Naive Bayes before using k-means clustering

used in the experiments. The used of statistical testing for the significance of the study and the evaluation metrics for the accuracy and error (misclassification) rate. The programming tool used to implement the algorithms is MATLAB. This is because MATLAB is a very powerful computing system for handling calculations involved in scientific and engineering problems. With MATLAB, computational and graphical tools to solve relatively complex science and engineering problems can be designed, developed and implemented. Specifically, MATLAB 2007b was used for the development. Table 2 illustrates how the false positive and false negatives are calculated. The first row displays the total non-spam. That is divided to True Negative (TN) and false positive. The second row is the same as a first row. This row contains the total spam in data sets. Some of them are dedicated wrongly as non-spam and the rest of them correctly dedicated as spam.

We provide two general testing parameters or function measurement that is commonly used in spam detection as following see Eq. (2) to (6):

$$\text{Accuracy rate} = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad (2)$$

$$\text{Error rate} = \frac{FP+FN}{TP+TN+FP+FN} * 100 \quad (3)$$

$$FP = \frac{FP}{FP+TP} * 100 \quad (4)$$

$$FN = \frac{FN}{FN+TN} * 100 \quad (5)$$

$$F - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Experimental results and analysis: In our proposed mechanism, these sections discuss the results. Figure 6 and Table 3 illustrate the gain charts and table for both training and testing results of the Naive Bayes before

Table 2: Illustrates how the false positive and false negatives are calculated

	Non-spam	Spam
Non-spam	True Negative (TN)	False Positive (FP)
Spam	False Negative (FN)	True Positive (TP)

Table 3: Analyzing of training and testing result using Naive Bayes before using clustering

Classification	Training	Training (%)	Testing	Testing (%)
Correct	2428	75.92	1029	73.34
Wrong	770	24.08	374	26.66
Total	3198		1403	

Table 4: Analysis of training and testing result using Naive Bayesian and K-means clustering

Classification	3167	99%	1354	96.5%
Correct	31	1%	49	3.5%
Wrong	3198		1403	
Total	3167	99%	1354	96.5%

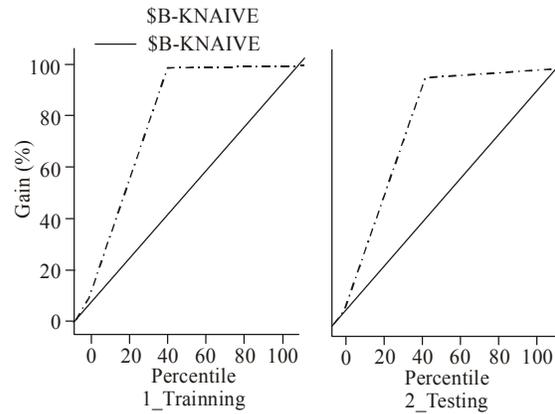


Fig. 7: Training and testing result without the best line (include baseline)

using a K-Mean clustering. Classification using Naive Bayes for training is at its best 75.92% and testing 73.34% accuracy. Gains chart with a baseline, the best line is (\$Best-Naive) and the result of Naive Bayes before improvement is (\$B-Naive).

Table 3 illustrates the results of the experiment for training and testing using Naive Bayes. The correct classification using Naive Bayes for training is at its best 75.92% and for testing 73.34% of accuracy while the wrong classification is 24.08% for training and 26.66% for testing.

Figure 7 illustrates the gain charts both training and testing results without the best line. To verify the best accuracy it must turn result equivalent to blue color line. Gains chart with a baseline, best line is (\$B-K Naive).

Figure 8 and Table 4 illustrate the gain charts and table for both training and testing results using hybrid between Naive Bayes and K-Mean clustering. To verify the best accuracy it must turn result equivalent to the blue line. The result after using hybrid of Naive Bayes and K-means clustering is verifying the best line that is

Table 5: Summary of results

Classifier	Accuracy of training (%)	Accuracy of testing (%)
Naïve Bayes	75.92	73.34
Naïve Bayes and K-means clustering	99	96.50

Table 6: Comparisons between different methods

Paper authors	Method	Accuracy of training (%)	Accuracy of testing (%)
Zhao and Zang (2005)	RSC	97.37	97.37
Taninpong and Ngamsuriyaroj (2009)	Multinomial NB	91.75	93.53
Salehi and Selamat (2011)	PSO and SAIS	88.33	88.33
Ismaila and Ali (2012)	SVM & ANN	97.78	97.26
Before improved	Naïve Bayes	75.92	73.34
After improved	Naïve Bayes and K-means clustering	99	96.50

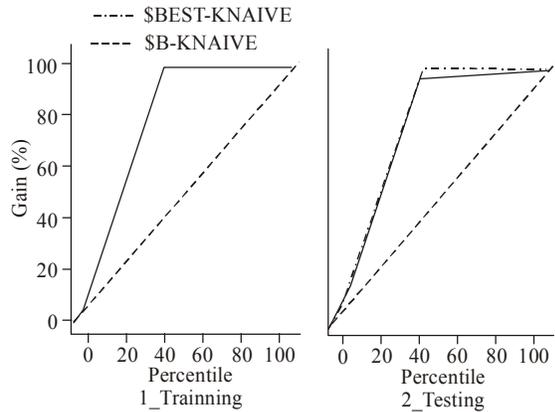


Fig. 8: Training and testing result for naïve after using k-means clustering (include best line)

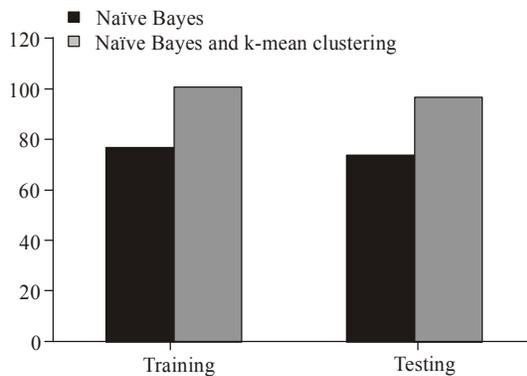


Fig. 9: Column accuracy comparisons between Naïve Bayes and hybrid between Naïve Bayes and k-means clustering

equivalent to the blue line. Classification using hybrid between Naïve Bayes and K-means clustering for training is at its best 99% and for testing 96.5% of accuracy while Naïve Bayes only is 75.92% for training and testing 73.34% accuracy. The final results are after using K-means clustering and Naïve Bayes are improving the accuracy and give a better result than the others. Gains chart with baseline and the result of Naïve Bayes after improvement is (\$B-KNAIVE).

Table 4 shows the result of the experiment for training and testing using the K-means clustering and Naïve Bayes. The correct classification using K-means

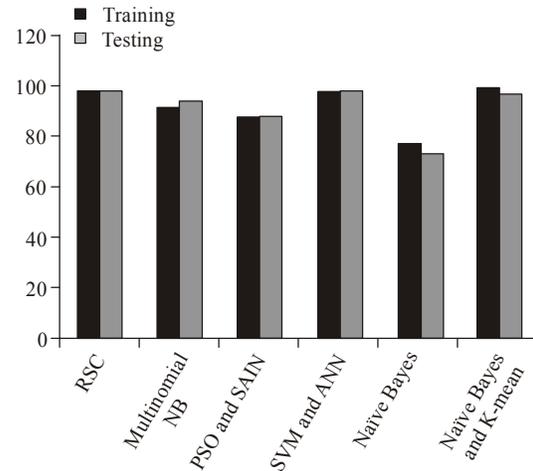


Fig. 10: Column accuracy comparisons between different methods to improved the accuracy

clustering and Naïve Bayes for training is at its best 99% and testing 96.5% accuracy while the wrong classification is 1% for training and 3.5% for testing. Table 5 illustrates the summarized results obtained after Naïve Bayes enhanced, while Fig. 9 shows the column of accuracy comparisons between Naïve Bayes and hybrid of (Naïve Bayes and K-Mean clustering). When comparing the result of using Naïve Bayes with the result of using the hybrid of (K-means clustering and Naïve Bayes) from accuracy perspectives, we find that the result of using the hybrid of the K-means clustering and Naïve Bayes are better than using Naïve Bayes.

Compression with other methods: This section shows the results of comparison between several different methods using enhanced spam detection. When comparing the result of different approaches with our result we find our result is better. Table 6 shows the summarized results obtained after comparing the result with others method. Figure 10 and 11 presents the accuracy comparisons between different methods.

From Table 5, Fig. 10 and 11, different methods were used to enhance spam detection and to reduce the misclassification rate (error rate). Zhao and Zang (2005) used Rough Set Classifier (RSC), in their study the result with divided the data set into 66.67 and 33.33% for training and testing respectively, for

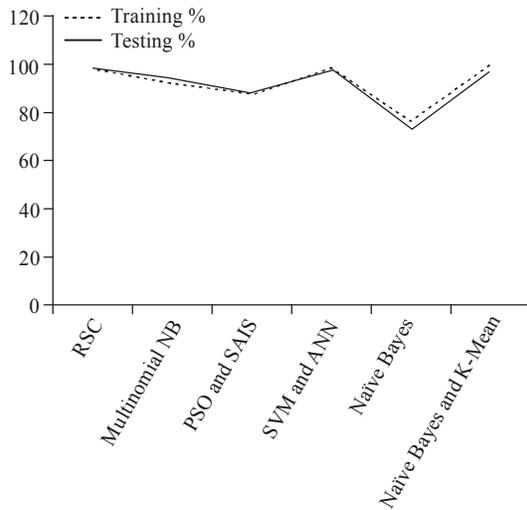


Fig. 11: Line accuracy comparisons between different methods

training was 97.37% of accuracy and the result of testing was 97.37 %. While the result with divided the data set into 80% and 20% for training and testing respectively, for training was 92.07% of accuracy and the result of testing was 92.07 %. Taninpong and Ngamsuriyaroj (2009) suggest an incremental adaptive spam mail classification based on multinomial Naïve Bayes classification and computed probability percentages to select the best feature to handle the problem of new spam and to reduce the misclassification percentage. The testing and training result using Trec05-1 is 91.75 and 93.53 of accuracy respectively.

Salehi and Selamat (2011) used Particle Swarm Optimization (PSO) and Simple Artificial Immune System (SAIS). They found that the training result was 88.33% while the testing result was 88.33%. Ismaila and Ali (2012) used Support Vector Machine (SVM) Artificial Neural Network (ANN), the result of training is 97.78% and testing is 97.26%. In the same study the result of Naïve Bayes (NB) is 75.92% for training and 73.34% for testing. This study enhanced Naïve Bayes by using K-means clustering; we found that the result of hybrid proposed Naïve Bayes and K-s clustering was 99% for training and 96.5% for testing data. The result of proposed method outperforming the previous spam detection mechanism mentioned above.

The Contributions of this study: A final result after using K-means clustering and Naïve Bayes improved accuracy and gives the best result than others. The advantages and contributions of this study can be summarized by these points.

The new spam detection mechanism is based on the combination between the Naïve Bayes and the K-means clustering techniques which none of the previous studies introduced:

- The combined method achieved better results in terms of classification accuracy on the spambase dataset.
- Our proposed method obtained good results by reducing the spam miss-classified rate compared with current spam detection methods such as Taninpong and Ngamsuriyaroj (2009), Zhao and Zang (2005).
- Enhancement of accuracy score by combing the K-means clustering method to the Naïve Bayes and proved that via correlation coefficient and T-test significance test algorithm.

STATISTICAL TESTING

In this study two statistical techniques were used: t-test and correlation coefficient for significance testing for accuracy prospective:

T-test for significant: T-test is the statistical technique used to select words that are more likely to be spam or non-spam. The formula of the T-test is shown as following:

$$Z_0 = \frac{|\bar{X}_1 - \bar{X}_s|}{\sqrt{\frac{\sigma_1^2 + \sigma_s^2}{n_1 + n_s}}}$$

From the formula,

\bar{X}_1 = The mean of non-spam mail

\bar{X}_s = The mean of spam mail

σ_1 = The variance of non-spam mail

σ_s = The variance of spam mail

n_1 = The number of non-spam mail

n_s = The number of spam mail

From Fig. 12, x-axis describes the frequency of each word in each email and y-axis describes the frequency of email that has this word. For example, if we have the frequencies of occurrence of the word “all” in spam and non-spam mail as 10 and 12 respectively, it is likely that the word “all” is not playing as a main feature in characterizing spam mail and non-spam mail. On the other hand, if we get the frequencies equal to 2 and 10 for the word “make” in spam mail and non-spam mail respectively, it is likely that the word “make” is significant in determining spam and non-spam mail. Thus, the word “make” should be selected as one of the features to classify spam mails. T-test statistical technique is the method that can be used to separate non-spam mail and spam mail when the difference between averages of spam and non-spam mail is high and the difference of variance is low. From Fig. 12, the left-hand side shows the low z_0 that graph has overlap area, so we cannot separate between spam and non-spam emails. The right-hand side of the figure shows the high z_0 with high contrast between spam and non-spam mails.

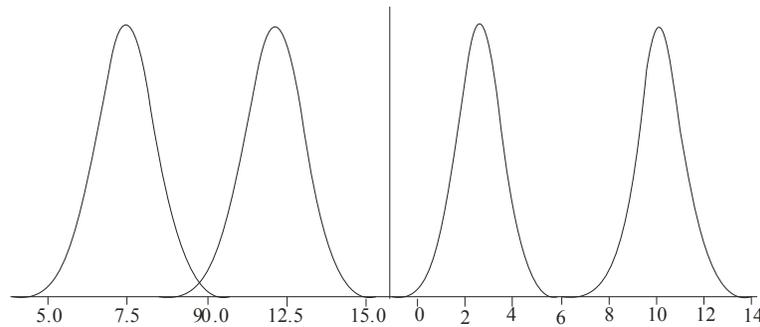


Fig. 12: t-test statistics, z_0

Table 7: Statistical significant testing using t-test

	Paired differences			95% confidence interval of the difference		
	Mean	S.D.	S.E. mean	Lower	Upper	Sig. (2-tailed)
Pair 1 Naïve-KNaive	-0.2287	0.1583	2.348E-03	-0.2333	-0.2241	0.0265

Table 8: Shows that our proposed method is statistically significant

Correlations coefficient	Naïve	KNaive
Naïve Pearson correlation	1.000	0.118
Sig. (2-tailed)	0	0.000
N	4601	4601
K Naïve Pearson correlation	0.118	1.000
Sig. (2-tailed)	0.000	0
N	4601	4601

Correlation coefficient is significant at the 0.01 level

In this study, t-test is used to show the significance size of our proposed method. Table 7 shows that our proposed method is statistically significant.

Table 7 displays the number of cases, mean value, standard deviation and standard error and significance values for the pairs of variables before and after the optimization use k-Mean clustering (Naive, KNaive) compared in the Paired Samples t-test method. The Paired-Samples t-test methods evaluate the means of two variables that stand for the same group at different times. The average values for the two variables (Naive, KNaive) are shown in the Paired Samples Statistics in Table 7. Since the Paired Samples t-test evaluate the means of the two variables, it is expedient to know what the mean values are. A low significance value for the t-test (typically less than 0.05) indicates that there is significance between the two variables. We can see in Table 7, the terms Sig. is (0.0265), which shows that our proposed method obtained significant results in Naive and KNaive. In addition, the significance value is high in Naive and KNaive values and the confidence interval for the mean difference does not contain zero. We can therefore conclude that there is a significant difference between results before and after optimization.

Correlation coefficient: Correlation Coefficient (CC) is a statistical tool used to study correlations between a set of variables. For example, a CC is used to learn a relationship between two variables and then the user

can take decisions on these relationships. Pearson's, Kendall and Spearman's correlation coefficients are well known CC types (Mohammed *et al.*, 2010). In this study, we used the Pearson's-CC in order to assess the efficiency of our proposed method.

Karl Pearson (Pearson, 1920) proposed Pearson's correlation coefficient. It gauges the linear relationship between two variables and the relationship is located between -1 and 1. The CC obtains the highest value i.e., +1 if the variable is growing in relationship and on other hand CC obtains -1 if variable is decreasing in relationship.

In this study, Pearson's correlation coefficient (r, ranging from -1 to 1) is used to show the significant size of our proposed method between Naïve Bayes before and after improvement. Table 8 shows that our proposed method is statistically significant, see equation below:

$$r = \frac{N \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{(N \sum_i x_i^2 - (\sum_i x_i)^2) (N \sum_i y_i^2 - (\sum_i y_i)^2)}}$$

According to correlation coefficient formula mentioned above, x_i is the Naïve Bayes value, \bar{x}_1 is the mean of Naïve Bayes value, y_i is the Naïve Bayes value after improved (K Naive), \bar{y}_1 is the mean of Naïve Bayes value after improved (KNaive).

Table 8 reflects comparison among original Naïve Bayes and our improved Naïve Bayes algorithm. According to Table 8 result Naïve Bayes is improved because the correlation coefficient result is less than the value of 0.01.

CONCLUSION

This study proposed a new scheme using a hybrid of the Naïve Bayes and K-means clustering for spam detection. In general, the proposed scheme offers the

following advantages such as improved classification accuracy and enhanced Naïve Bayes algorithm. The classification accuracy is 99% for training and 96.5% for testing. The result of the new mechanism using hybrid of Naïve Bayes and K-means clustering was compared with spam detection using Naïve Bayes only, it gives better classification accuracy and enhanced Naïve Bayes algorithm. There are adopted methods for ranking the 57 attributes to determine the most effective spam predictors, such as t-test statistic measure of significance to the difference between two means of each attribute for the spam and non-spam subsets of the data. Performance was compared with different methods namely artificial neural networks and negative selection algorithm classifiers reported in the literature for the same email dataset. The above comparison with different methods indicates that the new mechanism using hybrid of K-means clustering and Naïve Bayes provides better classification accuracy.

ACKNOWLEDGMENT

This study was financially supported in part by IDF in Universiti Teknologi Malaysia. The authors would like to thank the Research Management Centre (RMC) Universiti Teknologi Malaysia and Algraf Technical College-Sudan for their supports.

REFERENCES

- Alguliev, R.M., R.M. Aliguliyev and S.A. Nazirova, 2011. Classification of textual e-mail spam using data mining techniques. *Appl. Comput. Intell. Soft Comput.*, 2011 Article ID 416308, pp: 8.
- Arun, R. and T. Durga, 2009. Adaptive spam filtering based on bayesian algorithm. *Proceedings of 23rd International Conference on Computer Science, Information and Technology*, Pune, ISBN-978-93-81693-83-4.
- Attri, U. and H. Kaur, 2012. Comparative study of gaussian and nearest mean classifiers for filtering spam e-mails. *Int. J. Comput. Sci. Appl. (TIJCSA)*, 3(5): 2079-8407.
- Carpinter, J. and R. Hunt, 2006. Tightening the net: A review of current and next generation spam filtering tools. *Comput. Secur.*, 25(8): 566-578.
- Caruana, G. and M. Li, 2012. A survey of emerging approaches to spam filtering. *ACM Comput. Surv. (CSUR)*, 44(2): 9.
- Chhabra, P., R.Wadhvani and S. Shukla, 2010. Spam Filtering using Support Vector Machine. *Special Issue IJCTT*, 1(2, 3, 4): 166-171.
- Çiltık, A. and T. Güngör, 2008. Time-efficient spam e-mail filtering using n-gram models. *Pattern Recogn. Lett.*, 29(1): 19-33.
- Cournane, A. and R. Hunt, 2004. An analysis of the tools used for the generation and prevention of spam. *Comput. Amp. Secur.*, 23(2): 154-166.
- DeBarr, D. and H. Wechsler, 2009. Spam detection using clustering, random forests and active learning. *Proceeding of the 6th Conference on Email and Anti-Spam*, Mountain View, California, USA.
- Fagbola, T., S. Olabiyisi and A. Adigun, 2012. Hybrid GA-SVM for efficient feature selection in e-mail classification. *Comput. Eng. Intelli. Syst.*, 3(3): 17-28.
- Guzella, T.S. and W.M. Caminhas, 2009. A review of machine learning approaches to Spam filtering. *Expert Syst. Appl.*, 36(7): 10206-10222.
- Hayati, P. and V. Potdar, 2008a. Evaluation of spam detection and prevention frameworks for email and image spam: A state of art. *Proceedings of the 10th International Conference on Information Integration and Web-based Applications and Services*, pp: 520-527.
- Hayati, P. and V. Potdar, 2008b. Evaluation of spam detection and prevention frameworks for email and image spam: A state of art. *Proceedings of the 10th International Conference on Information Integration and Web-based Applications and Services*, pp: 520-527.
- Hershkop, S. and S.J. Stolfo, 2005. Combining email models for false positive reduction. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp: 98-107.
- Hong, C., 2011. Improving classification in Bayesian networks using structural learning. *World Acad. Sci. Eng. Technol.*, 75: 1407-1411.
- Ismaila, I. and S. Ali, 2012. Optimized spam classification approach with negative selection algorithm. *J. Theor. Appl. Inform. Technol.*, 39(1): 22-31.
- Lai, C.C. and C.H. Wu, 2007. Particle swarm optimization-aided feature selection for spam email classification. *Proceedings of the 2nd International Conference on Innovative Computing, Information and Control*, pp: 165.
- Long, X., W. L. Cleveland and Y.L. Yao, 2011. *Methods and Systems for Identifying and Localizing Objects based on Features of the Objects that are Mapped to a Vector: Google Patents*.
- Ma, W., D. Tran and D. Sharma, 2009. A novel spam email detection system based on negative selection. *Proceeding of the 4th International Conference on Computer Sciences and Convergence Information Technology, ICCIT '09*, pp: 987-992.
- MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *Proceeding of the 5th Berkeley Symposium on Math Statist and Prob.*, pp: 281-297.

- Manjusha, K. and R. Kumar, 2010. Spam mail classification using combined approach of bayesian and neural network. *Proceeding of the International Conference on Computational Intelligence and Communication Networks (CICN)*, pp: 145-149.
- MarkHopkins, E.R., G. Forman and J. Suermondt, 1999. Retrieved from: <http://archive.ics.uci.edu/ml/datasets/Spambase>.
- Marsono, M.N., M.W. El-Kharashi and F. Gebali, 2009. A spam rejection scheme during SMTP sessions based on layer-3 e-mail classification. *J. Network Comput. Appl.*, 32(1): 236-257.
- Mohammad, A.H. and R.A. Zitar, 2011. Application of genetic optimized artificial immune system and neural networks in spam detection. *Appl. Soft Comput.*, 11(4): 3827-3845.
- Mohammed, M., A. Shawkat and T. Kevin, 2010. Improved C4.5 algorithm for rule based classification. *Proceedings of 9th Artificial Intelligence, Knowledge Engineering and Database Conference (AIKED'10)*, pp: 296-301.
- Münz, G., S. Li and G. Carle, 2007. Traffic Anomaly Detection using K-Means Clustering. In *GI/ITG Workshop MMBnet, 2007*.
- Pearson, K., 1920. Notes on the history of correlation. *Biometrika*, 13(1), 25-45.
- Pour, A.N., R. Kholghi and S.B. Roudsari, 2012. Minimizing the time of spam mail detection by relocating filtering system to the sender mail server. *Int. J. Network Secur. Appl.*, 4(2): 10.
- Radicati, S. and Q. Hoang, 2011. Email Statistics Report, 2011-2015. Retrieved May, 25, 2011.
- Rao, I.K.R., 2003. Data mining and clustering techniques. *Proceeding of the DRTC workshop on Semantic Web*.
- Raskar, S.S. and D.Thakore, 2011. Text mining and clustering analysis. *IJCSNS*, 11(6): 203.
- Saad, O., A. Darwish and R. Faraj, 2012. A survey of machine learning techniques for Spam filtering. *Int. J. Comput. Sci. Network Secur.*, 12(2): 66.
- Sadan, Z. and D.G. Schwartz, 2011. Social network analysis of web links to eliminate false positives in collaborative anti-spam systems. *J. Network Comput. Appl.*, 34(5): 1717-1723.
- Sahami, M., S. Dumais, D. Heckerman and E. Horvitz, 1998. A Bayesian Approach to Filtering Junk E-Mail. *Learning for Text Categorization: Paper from the 1998 Workshop AAAI Technical Report WS-98-05*.
- Salcedo-Campos, F., J. Díaz-Verdejo and P. García-Teodoro, 2012. Segmental parameterisation and statistical modelling of e-mail headers for spam detection. *Inform. Sci.*, 195(0): 45-61.
- Salehi, S. and A. Selamat, 2011. Hybrid Simple Artificial Immune System (SAIS) and Particle Swarm Optimization (PSO) for spam detection. *Proceeding of the 5th Malaysian Conference in Software Engineering (MySEC)*, pp: 124-129.
- Sanz, E.P., J.M. Gómez Hidalgo and J.C. Cortizo Pérez, 2008. Email spam filtering. *Adv Comput.*, 74: 45-114.
- Schneider, K.M., 2003. A comparison of event models for Naive Bayes anti-spam e-mail filtering. *Proceeding of the 10th Conference of the European Chapter of the ACL, Buda-Pest, Hungry*, pp: 307-314.
- Schryen, G., 2007. The impact that placing email addresses on the Internet has on the receipt of spam: An empirical analysis. *Comput. Secur.*, 26(5): 361-372.
- Tala Tafazzoli, S.H.S., 2009. A combined method for detecting spam machines on a target network. *Int. J. Comput. Networks Commun. (IJCNC)*, 1(2): 35-44.
- Taninpong, P. and S. Ngamsuriyaroj, 2009. Incremental adaptive spam mail filtering using naïve bayesian classification. *Proceeding of the 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing*, pp: 243-248.
- Torres, G.J., R.B. Basnet, A.H. Sung, S. Mukkamala and B.M. Ribeiro, 2009. A similarity measure for clustering and its applications. *Int. J. Electric. Comput. Syst. Eng.*, 3(3).
- Vinther, M., 2002. Intelligent Junk Mail Detection Using Neural Networks. Retrieved form: URL: <http://www.logicnet.dk/reports/JunkDetection/JunkDetection.pdf>.
- Vorakulpipat, C., V. Visoottiviseth and S. Siwamogsatham, 2012. Polite sender: A resource-saving spam email countermeasure based on sender responsibilities and recipient justifications. *Comput. Amp. Secur.*, 31(3): 286-298.
- Wu, H., H. Li, G. Wang, H. Chen and X. Li, 2011. A novel spam filtering framework based on fuzzy adaptive particle swarm optimization. *Proceeding of the International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pp: 38-41.
- Xiao-Li, C., L. Pei-Yu, Z. Zhen-Fang and Q. Ye, 2009. A method of spam filtering based on weighted support vector machines. *Proceeding of the IEEE International Symposium on IT in Medicine and Education*, pp: 947-950.
- Yin, H. and Z. Chaoyang, 2011. An improved bayesian algorithm for filtering spam e-mail. *Proceeding of the 2nd International Symposium on Intelligence Information Processing and Trusted Computing (IPTC)*, pp: 87-90.
- Ying, K.C., S.W. Lin, Z.J. Lee and Y.T. Lin, 2010. An ensemble approach applied to classify spam e-mails. *Expert Syst. Appl.*, 37(3): 2197-2201.
- Youn, S. and D. McLeod, 2007. A comparative study for email classification. *Adv. Innovat. Syst. Comput. Sci. Software Eng.*, pp: 387-391.

- Zhang, H. and D. Li, 2007. Naïve Bayes Text Classifier. pp: 708-708.
- Zhang, Q., H. Yang, P. Wang and W. Ma, 2011. Fuzzy clustering based on semantic body and its application in chinese spam filtering. *JDCTA: Int. J. Dig. Content Technol. Appl.*, 5(4): 1-11.
- Zhao, W. and Z. Zhang, 2005. An email classification model based on rough set theory. *Proceedings of the 2005 International Conference on Active Media Technology, (AMT 2005)*, pp: 403-408.