Feature Weight Optimization Mechanism for Email Spam Detection based on Two-Step Clustering Algorithm and Logistic Regression Method

Ahmed Hamza Osman
Department of Information System,
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

Hani Moaiteq Aljahdali
Department of Information System,
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

Abstract—**This research proposed an improved filtering spam technique for suspected emails, messages based on feature weight and the combination of two-step clustering and logistic regression algorithm. Unique, important features are used as the optimum input for a hybrid proposed approach. This study adopted a spam detector model based on distance measure and threshold value. The aim of this model was to study and select distinct features for email filtering using feature weight method as dimension reduction. Two-step clustering algorithm was used to generate a new feature called "Label" to cluster and differentiate the diversity emails and group them based on the inter samples similarity. Thereby the spam filtering process was simplified using the Logistic regression classifier in order to distinguish the hidden patterns of spam and non-spam emails. Experimental design was conducted based on the UCI spam dataset. The outcome of the findings shows that the results of the email filtering are promising compared to other modern spam filtering methods.**

*Keywords—**Two-step clustering; spam filtering; classification; detection; feature weight; logistic regression***

I. Introduction

Nowadays, email messages are considered as economic and most essential communicative way in the world. It is efficient, simple and accessible for all due to the internet availability. The availability of email makes it susceptible to many hackers and threats [1]. Spam is considered as a very important threat to email; practically all email users in the world tolerate spam. The term spam was used to define the undesirable message, junk-mails sent to web users' inbox. It is most opportune for email spammers to send lots of messages to millions of users simply and without cost [2]. This makes it a public situation for all web users to receive unsolicited email regularly.

The versatile way of unsolicited email by the utilization of immense mailing tools prompts the requirement for spam recognition. Execution of various spam discovery strategies in view of machine learning methods was proposed to address the issue of various email spam desolating the system. Past calculation utilized as a part of email spam identification contrasts each email message and spam and non-spam information before creating finders. This study' proposed system propelled by the two-step grouping calculation with strategic relapse system utilizes highlights weight as advancement procedure to produce locators to cover the spam space.

Diverse strategies have been embraced to stop the danger of spam or to definitely lessen its measure. An anti-spam law was authorized by enacting a punishment for spammers who circulate email spam [3]. In spite of the diverse methodologies and strategies that have been received to battle the danger of email spam, the web today still shows a huge measure of spam [4]-[6]. Therefore, more consideration is required with respect to how the risk can be radically diminished if not completely disposed. The fight against email spam is an extremely troublesome fight; therefore, it bodes well to battle a versatile email spam generator with a versatile system.

In this study, a new hybrid method that is inspired by descriptive and predictive models will be introduced. It consists of a Logistic Regression Method (LRM) as a prediction method with the integrated effort of Two-step Clustering Algorithm (TSCA) as description technique. To produce more precise filtering results, the standard dimension of spam dataset has been reduced based on feature weight (FW). The engineering aims required in this study's hybrid method can be viewed in three ways; firstly, generating new dataset based on feature weight (FW) to reduce the dataset dimensionality; secondly, to limit the maximizing distance between spam detectors and the non-spam space by using two-step clustering algorithm (TSCA); and thirdly, is to filter the email to spam and no-spam using logistic regression method (LRM) based on the output of FW and TSCA. The aim of this study is to find possible increase in the accuracy and reduction in the miss-filtering emails.

This article is structured into six sections: Section 1 discusses the motivation and Introduction; Section 2 covers the article related work, the improved method, and its integral system will be described in Section 3. Experimental design and results of the study and discussions in details are in Section 4 and Section 5, respectively. The conclusion of the research is described in Section 6.

II. Related Works

Several attempts have been proposed to block spammers and reduce a number of undesirable emails across the internet and user's inbox. One of these attempts is called anti-spam law [3]. This law was defined by enacting a penalty for spam users who send spam emails to user's inbox. Another two common methods have been proposed in email spam detection; a Machine Learning (ML) method, a data mining (DM) and knowledge discovery (KDD) method [4]. In the DM method, researchers introduced an origin-based filter technique based on web protocol address approach to differentiate the spam and non-spam messages. On the other hand, in the KDD method, researchers categorized spam or non-spam message based on sets of generating rules using KDD algorithms as filter techniques. The authors claim a promising spam filtering results. However, they need to update the rules continuously, which is time wasting and inadequate for many users. Spam detection based on ML is not required to generate and update any rules as DM and KDD based methods; only training data for classifying an email message is required. Classification techniques based on email messages characteristics were applied to learn the filtering rules and to distinct spam and non-spam email messages [5].

Some approaches were adopted to stop the spam, however, the web still currently observe a large set of spam [6], [7]. Therefore, more consideration is required by improving spam detection algorithm on how the threat can be significantly decreased if not completely excluded. For this aspect, many spam-filtering algorithms have been applied in machine learning [5]. Examples of these algorithms include neural network (NN), Support Vector Machine (SVM), k-nearest neighbor (KNN), and Naïve Bayes (NB). Several studies in machine learning approach applied in email spam filtering (Table 1). Marsono et al. [8] implemented naïve Bayes email spam filtering based on layer processing, without any requirement for reassembling. They suggested controlling middle boxes step to filter the received email spam from the email servers [9]. W. El-Kharashi et al. proposed a spam controlling method using hardware structure of naïve Bayesian inference engine [10]. The method can categorize more than 117 million features per-second based on probability inputs [10]. Y. Tang et al. introduced a model that applied the SVM for email filtering. This model extracts spammers behavior using the distribution of the global senders and then investigate them by assigning a value of no-spam to each IP-address email sender [11]. Their empirical results presented that the SVM technique is precise and faster than the Random Forests (RF) algorithm [11]. Yoo, S., et al. presented an email classification method called Priority E-mail Personalized technique (PEP) [12]. The PEP focused on analyzing the personal social networks to detect user groups and to achieve the user viewpoint based on the user social roles and then apply them for email message classification. Silva et al. [13], [14] assessed the neural network algorithm for internet spam. They also investigated how different groups of features influence the filtering accuracy rate. Largilliere and Peyronnet [15] developed a combination approach for internet email spamming on the PageRank method. Liu et al. [16] introduced features of user behavior for distinguishing spam and non-spam pages. They also developed a hybrid machine learning system aided by user-behavior to filter spam pages [16]. Content-based features method were proposed by Castilho et al. [17] and Rungsawang et al. [18]. These studies investigated and extracted both link features and content for spam filtering pages with some improving email spam detection using ant colony optimization method [18]. Also, they used the topology of the web-graph by extracting the web link dependencies between the internet pages.

The logistic regression method has some benefits compared to other classification methods such as SVM and Naive Bayes. The excessively robust conditional independence assumptions of Naïve-Bayes and SVM mean that if two variables are correlated, the naïve-Bayes and SVM will multiply them together as if they were independent, overrating the evidence. On the other hand, the LR is much more strong to correlated variables; if two features (A) and (B) are faultlessly correlated, LR will only allocate half the weight to w(A) and a half to w(B). Thus, when there are various correlated variables, LR will simply allocate a more precise probability than the SVM and naïve-Bayes. This LR is better than many other data mining methods in the small and large dataset [19], [20]. These reasons prompted the investigation and examination of the LR in spam email filtering.

TABLE. I. Spam Detection based on ML

Study	ML algorithms	Advantages
Marsono et al. [9]	Naïve Bayes	Filtering spam based on layer processing, without any requirement for reassembling.
W. El-Kharashi et al. [10]	Naïve Bayesian inference engine	Categorizing more than 117 million features per-second based on probability inputs
Y. Tang et al. [11]	SVM	SVM technique is precise and faster than the Random Forests (RF) algorithm
Silva et al. [13]	Neural network algorithm	Investigating how different groups of features influence the filtering accuracy rate
Yoo, S., et al. [12]	Priority E-mail Personalized technique (PEP)	Analyzing the personal social networks to detect user groups and to achieve the user viewpoint based on the user social roles and then applying them for email message classification
Largilliere and Peyronnet [15]	Combination approach for internet email spamming	PageRank method
Liu et al. [16]	A hybrid machine learning system aided by user-behavior to filter spam pages	Features of user behavior for distinguishing spam and non-spam pages.
Castilho et al. [17] & Rungsawang et al. [18]	Content-based features method	Extracting both link features and content for spam filtering pages with some improving email spam detection using ant colony optimization method and the topology of the web-graph

III. Proposed Model and Operational System

The presented improved model and its constituent systems upgraded strategies in current circumstances have broad achievement in numerous true complex critical thinking. The significance of a joint system is not debatable, in light of the way that an individual system has its shortcoming, and an enhanced system is intended to complement the shortcoming of these individual shrewd systems. A brilliant mix of two-step bunching calculation and strategic relapse strategy is researched keeping in mind the end goal to compliment the parameters of every segment of the system. This is work by utilizing the benefits of an individual system against its inconveniences while lifting each powerless segment individual from both systems to accomplish dependability, consistency and a precise keen system extendable for utilization in grouping. The proposed enhanced system is utilized to shape a superior enhanced system with weighted elements in light of highlight weight handle.

This proposed method combined with different techniques such as Two-step clustering algorithm and logistic regression. The integrated techniques are then applied through several steps such as pre-processing (dividing the dataset into training and testing data) and weighing each feature based on the average values that can generate from each feature. The proposed system model is demonstrated in Fig. 1.

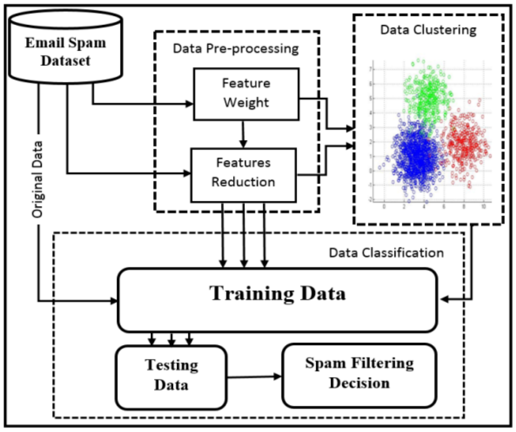


Fig. 1. Proposed system model.

A. Data Pre-processing

Pre-processing is one of the important data mining steps to prepare the dataset before the mining procedure. In this study, data preparation was used (and the dataset were divided into training and testing part), feature weight and feature reduction were based on feature weight step as three initial phases in this stage.

For preparing the dataset, there are several benchmark datasets for email spam classification and clustering roles [21]. One of this dataset is called Spam based which was reported by UCI Machine Learning repository and used in the spam filtering research such as [22], [23]. The main function of this dataset is to test and classify email messages to spam and non-spam messages. The spam based data is collected of 4,601 e-mails messages with 39.4 % (1,813) messages marked as Spam and 60.6 % (2,788) reported as non-spam [24].

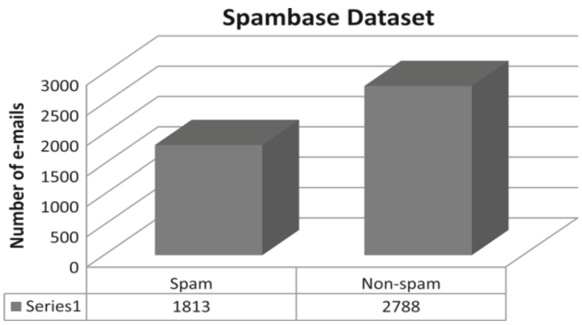


Fig. 2. Dataset distribution to spam and non-spam emails.

Fig. 2 shows the investigation of e-mail messages (spam and non-spam). In the proposed method using two-step clustering and logistic regression, the dataset was divided into 10 parts as 10-fold validation to examine the variation of the whole dataset. These parts employed for training and testing data. Each part consists of 460 instances except the last part, which consists of 461 instances. The proposed method was evaluated 10 times with each time nine parts employed as training dataset and one part considered as testing. In each round, it was considered that the testing part will be replaced with one of the nine training parts of the test and each part are done separately.

A combination of two-step clustering and logistic regression was conducted for training classifiers using the generated spam and non-spam features to filter the testing sample.

B. Data Clustering using Two-Step Algorithm

The two-step clustering technique is connected to wildcat algorithm developed to reveal natural groups inside a data set that might or not be clear [25]. The algorithm employed by this procedure has many captivating options that discriminate it from ancient clustering approaches:

- Ability to produce clusters in a continuous and categorical data type.
- The algorithm can control the generated clusters automatically.
- Ability to interact with a huge dataset probably.

C. Clustering Fundamental

The two-step technique uses distance criteria to handle continuous and categorical dataset. The likelihood considers that the data variables in the cluster system are freelance. Also, each categorical data is intended to own a multinomial distribution, and each continuous data is predictable to own a Gaussian distribution. Empirical interior testing determines that the procedure is efficiently strong to violations of each belief of independence and therefore the spatial arrangement assumptions. Conversely, it is necessary to try to remember that some of these assumptions are met. The two-steps of the technique's rule are summarized as follows:

- **First Step.** Pre-clustering the instances (or cases) into many small sub-groups. The procedure begins with the development of a Cluster Feature (CF) Tree. The tree starts by placing the first instance at the root in a leaf node that carries variable information for that instance. Every consecutive instance is then additional to associate present node or forms a new node according to the similarity between the current nodes.

- **Step 2.** Cluster the sub-groups resulting from pre-clustering step into the coveted number of groups. It can also choose the cluster number automatically. By using agglomerative clustering (AC) approach, the leaf nodes of the Cluster Features tree are then grouped. The AC can be conducted to range the produced solutions. The optimum number of clusters can be specified by comparing these clusters based on the Akaike Information Criterion (AIC) or Schwarz's Bayesian Criterion (BIC). The similarity scores between items calculated using an Euclidean distance measure that is described in (1).

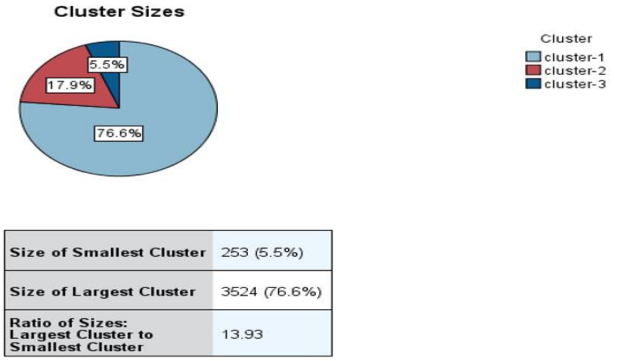
$$\text{Dist}\,(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

```
Algorithm Two-step Clustering
Input:
            //A Set X of object {$X_1,…X_n$}
            //A distance function dist($C_1$, $C_2$)
Output:
           //A Set X of clusters object {$X_1,…X_n$}
      For $i$ = 1 to $n$
           $C_i$={$X_i$}
      End for
      C={$C_1$,…,$C_n$}
        I=n+1
      While C.size > 1 do
              ($C_{min1}$, $C_{min2}$)= minimum dist($C_i$,$C_j$) for all $C_i$,$C_j$ in C
              Remove Cmin1 and Cmin2 from C
              Add {$C_{min1}$, $C_{min2}$} to C
              I = I+1
     End While
```

An Euclidean vector is the position of a point in a likelihood n-space. Therefore, X is (Xn, Xn, … , Xn) and Y is (Y1, Y2, … , Yn) are likelihood vectors, starting from the origin of the space, and two points are indicated by their tips [26]. The Two-step algorithm process is demonstrated as above.

The distribution of the email messages and clustering representation process using two-step clustering algorithm is demonstrated in Fig. 3.

Fig. 3 represents the clustering output using the two-step clustering method to cross the spam dataset. It was observed that the number of extracted clusters is 3. One of the advantages of the two-step clustering algorithm is that it has the ability to determine the number of clusters automatically. An observation was noted that the size of the small cluster is cluster 3 with 253 (5.5%) email messages distribution ratio. On the other hand, the largest cluster size is cluster 1 with 3524 (76.6). The ratio of cluster 1 to cluster 3 is 13.93%. A new feature labeled as cluster represents the output of these clusters. By this feature, we can integrate the clustering algorithm with another mining method for a possible improvement reason.



Size of Smallest Cluster	253 (5.5%)
Size of Largest Cluster	3524 (76.6%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	13.93

Fig. 3. Clustering results using the two-step clustering method.

D. Data Classification using Logistic Regression

Logistic regression is considered as one of the important statistical methods for investigating data in which there is one or more autonomous feature that defines results. The results are measured with a dichotomous feature, which means that the possible outcomes are two only. Based on the logistic regression mechanism, the dependent variable can be dichotomous or binary. For example, the data can only be coded as 1 (positive, Spam, Malware, detect, etc.) or 0 (negative, non-spam, non-malware, not detected, etc.). One of the main aims of the logistic regression is to find the optimum fitting model to represent the association between a set of predictor (independent) features and the interest dichotomous characteristic. Logistic regression extracts the significance levels and standard faults named coefficient values. The equation to classify a logic transformation probability of occurrence of the interested characteristic formulates as:

$$\text{odds} = \frac{\text{p}}{1 - \text{p}} = \frac{\text{Probability of presence of characteristic}}{\text{Probability of absence of characteristic}} \quad (2)$$

And

$$\text{logit}(\text{p}) = \ln\left(\frac{\text{p}}{1 - \text{p}}\right) \qquad (3)$$

In the classification based on logistic regression, only two classes y = 0, and y = 1 is formulated. A parametric form of P(y = 1 | x, w) is considered where w is the parameter vector.

$$\text{P}(\,\text{y} = 1 \mid \text{X}; \text{W}\,) = \text{P}_1(\text{X}) = \frac{1}{1 + \text{e}^{-\text{w.x}}} \qquad (4)$$

$$\text{P}(\,\text{y} = 0 \mid \text{X}; \text{W}\,) = 1 - \text{P}_1(\text{X}) \qquad (5)$$

It is informal to illustrate that this is equivalent to

$$\log = \frac{\text{P}(\,\text{y} = 1 \mid \text{X}; \text{W}\,)}{\text{P}(\,\text{y} = 0 \mid \text{X}; \text{W}\,)} = \text{W}.\text{X} \qquad (6)$$

The log odds of class 1 are a linear function of x as an example.

The proposed method used the discussed classifier using logistic regression to classify and filter the email into spam and non-spam. The experimental design based on the logistic regression will be discussed in the next section.

IV. Experimental Design

This experiment aimed to detect and filter the spam and non-spam messages from the email messages. The experiments were implemented on 4061 email messages, each message located as spam or non-spam according to the Spambase dataset. A method was executed by searching for the spam and non-spam email messages within the original dataset.

The spam dataset was broken down into 10 sets. Each set had a certain number of instances (email messages). The instances increased for each set with each weighting test round, starting with 460 email messages in the first set. Then, adding 460 more instances to the first set, and then, multiplying the amount of the data by 2, 3, 4, … 10 for the second set, third set, fourth set, to the tenth set, respectively. The objective of this grouping procedure was to study the pattern of the spammer user for each message so it can be focused. The average value of each of the features in the dataset was calculated as a first stage and it was noted that some of the features conveyed a very small value or had inverse proportion and some of them had a direct proportion between the number of instances and the feature values when the average was calculated. These pointers reflected the increasing and decreasing weighted score between the email features and the pattern of the spammer writing style. Possible hypothesis about this assumption was seen as a threshold for selecting the important features from unimportant features. The significant features were then nominated to enter the second training and testing experiment process. Conversely, the features that had a reverse proportion were ignored.

Training and Testing were implemented once again after features selection. The accuracy was declining as compared to the first experiment which caused the degree of learning depending on the number of significant features extracted from the email messages, and the decreasing of insignificant feature consequently led to the rise of the filtering accuracy and vice versa. The accuracy score was computed, and then the Spam base dataset was employed for training and testing process. The significant features that were selected based on the weighted process are shown in Table 2.

Table 2 demonstrates the sample results across the group of instances (messages). We have 57 features represented in each email message, and one feature named (class) represents the type of suspected message either spam or non-spam. According to the average values of these features, it was observed that several features conveyed a very small value or had inverse proportion. This score indicates that the feature is unimportant or not effectively on the filtering process of spam and non-spam. On the other hand, the significant features were reported in Table 2. This table represents features that had a direct proportion and definitely can affect the classification result by filtering the email messages to spam or non-spam. The weighting for each feature were computed to improve the achieved results that were obtained in Table 4 according to the following formula:

$$\mathrm{WF_i} = \frac{\sum \mathrm{F(i)}}{\mathrm{F(i)}} \qquad (7)$$

Where, WF_i = the weight of feature in the instance I; F(i)= Total number of values in feature i; i = (406, 920, 1380, 1840, 2300, 2760, 3220, 3680, 4140, and 4601). After the improvement process using feature weight, the effect of the weight enforcing the observation in inverse and direct proportion was observed.

TABLE. II. Significant Features

Feature ID	**Average Weighted Values**	**Feature Rank**	**Feature ID**	**Average Weighted Values**	**Feature Rank**
Feature 57	382.8014319	1	Feature 7	0.200511475	16
Feature 56	78.93747246	2	Feature 45	0.191798894	17
Feature 55	7.735299963	3	Feature 2	0.185727953	18
Feature 19	1.912533257	4	Feature 23	0.17626928	19
Feature 21	1.103635778	5	Feature 22	0.1670505	20
Feature 12	0.553360017	6	Feature 26	0.159371485	21
Feature 5	0.421804941	7	Feature 8	0.156010777	22
Feature 27	0.383728656	8	Feature 24	0.149887429	23
Feature 52	0.379102803	9	Feature 20	0.149717065	24
Feature 16	0.374121694	10	Feature 6	0.133869627	25
Feature 25	0.349365515	11	Feature 9	0.131386474	26
Feature 3	0.340623222	12	Feature 53	0.124596921	27
Feature 10	0.285855432	13	Feature 50	0.123808543	28
Feature 18	0.249953756	14	Feature 13	0.115471878	29
Feature 17	0.220250244	15	Feature 4	0.106688987	30

V. Results and Discussion

In this study, the experiments were built based on two types (original and weighted) spam datasets. The original dataset is the common spam data that was normally used in spam filtering research, while the weighted dataset is generated from the original dataset (Spambase) by calculating the average of each feature inside the original data. The reason for the weighted data is to study the pattern of the spammer for each feature and distinguish it as a significant or non-significant. Thus, the voted features that were selected based on the weighted process only can be used for spam filtering. By selecting the important features, the spam filtering performance will increase due to the features reduction that occurred by weighting process. To classify and filter the email messages, different types of an empirical study based on logistic regression and two-step clustering algorithm were conducted. The results that were generated behind the hypothesis will be presented in different phases: Logistic regression with all features in the dataset, logistic regression based on important features only, hybrid two-step and logistic regression with all Spam base feature datasets and the combined two-step with logistic regression based on important features that were extracted using feature weight process. The filtering accuracy computed based on the equation:

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FP}) + (\text{TP} + \text{FN})} \times 100 \qquad (8)$$

Where,

True Positive (TP): The number of spam and non-spam emails executable correctly classified; False Positive (FP): The number of spam executable classified as non-spam; True Negative (TN): The number of spam and non-spam executable incorrectly classified; False Negative (FN): The number of non-spam executable classified as spam emails.

The results of emails filtering using logistic regression methods based on dataset features and important features are illustrated in Tables 3 and 4, respectively.

The tables show the results of 10-fold cross validation to examine all the parts of the dataset. Each part implemented in one round from round 1 to round 10. For each experimental round, nine parts represent a training dataset while the remainder part (only one part) represents a testing dataset. The testing part is becoming one of the training datasets during each experiment. The total results are an equal average value for all the ten parts. These results represent the filtering accuracy of the training and testing data, the misfiltering ratio, the area under the carafe, and the number of correct filtering messages to spam and non-spam in the dataset.

TABLE. III. RESULT OF LOGISTIC REGRESSION WITH ALL FEATURES IN THE DATASET

Dataset Round	**Classification Accuracy TP/FN**		**Misclassification Accuracy TN / FP**		**Area under the Carafe**		**Number of corrected filtered email messages**			
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**		**Testing**	
							Spam	**Non-Spam**	**Spam**	**Non-Spam**
Round 1	90.51%	95.87%	9.49%	4.13%	0.962	0.991	3,748	393	441	19
Round 2	90.51%	94.78%	9.49%	5.22%	0.963	0.976	3,748	393	436	24
Round 3	90.80%	93.48%	9.20%	6.52%	0.962	0.981	3,760	381	430	30
Round 4	90.17%	95%	9.83%	5%	0.961	0.990	3,734	407	437	23
Round 5	91%	96.30%	9%	3.70%	0.962	0.993	3,749	392	443	17
Round 6	90.75%	94.57%	9.25%	5.43%	0.964	0.986	3,758	383	435	25
Round 7	90.22%	100%	9.78%	0.00%	0.959	1	3,736	405	460	0
Round 8	90.85%	95.43%	9.15%	4.57%	0.962	0.991	3,762	379	439	21
Round 9	91.23%	93.26%	8.77%	6.74%	0.966	0.976	3,778	363	429	31
Round1 0	92.44%	90.89%	7.56%	9.11%	0.971	0.968	3,827	313	419	42
Average	90.85%	94.96%	9.15%	5.04%	0.96.36	0.98.57	3760	380.9	436.9	23.2

TABLE. IV. RESULT OF LOGISTIC REGRESSION WITH IMPORTANT FEATURES

Dataset Round	**Classification Accuracy TP/FN**		**Misclassification Accuracy TN / FP**		**Area under the Carafe**		**Number of corrected filtered email messages**			
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**		**Testing**	
							Spam	**Non-Spam**	**Spam**	**Non-Spam**
Round 1	92.8%	90.22%	7.17%	9.78%	0.977	0.999	3,844	297	415	45
Round 2	92.97%	97.17%	7.03%	0.0283	0.976	0.990	3850	291	447	13
Round 3	93.19%	95.87%	6.81%	0.0413	0.976	0.995	3859	282	441	19
Round 4	92.71%	96.74%	7.29%	0.0326	0.974	0.997	3839	302	445	15
Round 5	93%	98.7%	7%	0.013	0.976	0.998	3851	290	454	6
Round 6	92.95%	96.96%	7.05%	0.0304	0.977	0.998	3849	292	446	14
Round 7	92.44%	99.35%	7.56%	0.0065	0.973	0.999	3828	313	457	3
Round 8	92.54%	99.57%	7.46%	0.0043	0.975	1	3832	309	458	2
Round 9	93.53%	94.57%	6.47%	0.0543	0.978	0.993	3873	268	435	25
Round1 0	94.15%	95.66%	5.85%	0.0434	0.981	0.992	3898	242	441	20
Average	93.03%	96.48%	6.97%	3.52%	0.97.68	0.9961	3852.3	288.6	443.9	16.2

In Table 2, it was observed that the achieved results on the 30 important features excluding the target feature (Class) are better than using all the dataset features. This indicates that the selected features are more significant. Also, the process time will be reduced accordingly because only the important features extracted will be tested rather than all features. Another criterion that was used for evaluating the proposed method is the Area under carafe (AUC). It is an assessment metric normally used in binary classification challenge. When the accuracy computed based on the true and false positive rate as the threshold rate for classifying an element as 0 or 1: if the predictor is best, the true positive ratio will rise rapidly, and the AUC will be close to 1. On the other hand, if the predictor is less than the random predicting, the true positive ratio will rise linearly with the false positive ratio and the AUC will be around 0.5 [27], [28]. AUC metric is important because it can evaluate the predictor's performance on the unbalanced dataset. It is independent of the fraction, of the test population, which is, target, class one, or zero. However, the spam and non-spam dataset that was used is not equivalent. The AUC results represented in Table 4 indicate that the performance evaluation is enforcing the filtering accuracy results and proved better results after weighting process and feature selection.

TABLE. V. RESULT OF HYBRID TWO-STEP AND LOGISTIC REGRESSION WITH ALL SPAM BASE FEATURES DATASET

Dataset Round	Classification Accuracy TP/FN		Misclassification Accuracy TN / FP		Area under the Carafe		Number of corrected filtered email messages			
	Training	Testing	Training	Testing	Training	Testing	Training		Testing	
							Spam	Non-Spam	Spam	Non-Spam
Round 1	97.33%	98.35%	2.67%	1.65%	0.986	0.989	4,072	69	457	3
Round 2	97.53%	95.96%	2.47%	4.04%	0.987	0.989	4,080	61	446	14
Round 3	97.12%	97.48%	2.88%	2.52%	0.986	0.989	4,063	78	453	7
Round 4	97.14%	98.57%	2.86%	1.43%	0.986	0.99	4,064	77	458	2
Round 5	97%	98.13%	3%	1.87%	0.986	0.989	4,070	71	456	4
Round 6	97.33%	96.39%	2.67%	3.61%	0.986	0.989	4,072	69	448	12
Round 7	97.14%	98%	2.86%	2.00%	0.986	0.989	4,064	77	455	5
Round 8	97.21%	96.83%	2.79%	3.17%	0.987	0.989	4,067	74	450	10
Round 9	97.26%	97.70%	2.74%	2.30%	0.986	0.989	4,069	72	454	6
Round1 0	97.53%	96.61%	2.47%	3.39%	0.987	0.989	4,079	61	450	11
Average	97.26%	97.40%	2.74%	2.60%	0.9863	0.9891	4070	70.9	452.7	7.4

TABLE. VI. RESULTS OF COMBINED TWO-STEP WITH LOGISTIC REGRESSION BASED ON IMPORTANT FEATURES

Dataset Round	Classification Accuracy TP/FN		Misclassification Accuracy TN / FP		Area under the Carafe		Number of corrected filtered email messages			
	Training	Testing	Training	Testing	Training	Testing	Training		Testing	
							Spam	Non-Spam	Spam	Non-Spam
Round 1	98.33%	99.35%	1.67%	0.65%	0.996	0.999	4,072	69	457	3
Round 2	98.53%	96.96%	1.47%	3.04%	0.997	0.999	4,080	61	446	14
Round 3	98.12%	98.48%	1.88%	1.52%	0.996	0.999	4,063	78	453	7
Round 4	98.33%	99.57%	1.67%	0.43%	0.996	1	4,073	68	458	2
Round 5	98%	99.13%	2%	0.87%	0.996	0.999	4,070	71	456	4
Round 6	98.33%	97.39%	1.67%	2.61%	0.996	0.999	4,072	69	448	12
Round 7	98.50%	98.26%	1.50%	1.74%	0.996	0.999	4,080	61	452	8
Round 8	98.55%	97.83%	1.45%	2.17%	0.997	0.999	4,082	59	450	10
Round 9	98.48%	98.48%	1.52%	1.52%	0.996	0.999	4,079	62	4,079	63
Round10	98.89%	98.26%	1.11%	1.74%	0.997	0.999	4,095	45	453	8
Average	98.41%	98.37%	1.59%	1.63%	0.999	0.999	4076.6	64.3	815.2	13.1

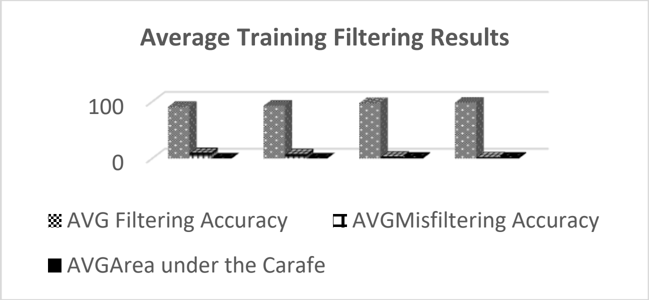


Fig. 4. Average training results of emails spam filtering experiments.

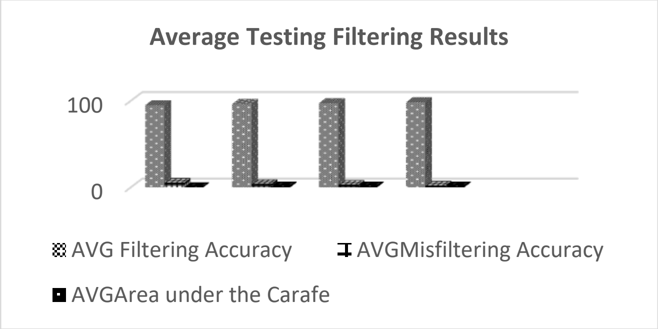


Fig. 5. Average testing results of emails spam filtering experiments.

Fig. 4 and 5 represent the average training output of the spam email filtering before and after feature selection using feature weight process. The dataset was examined based on two techniques; the logistic regression and the combined technique between logistic regression and two-step clustering algorithm. Table 3 presents the prediction of email filtering using the logistic regression method extracted average accuracy results with 90.8% for training phase and 94.96% for testing phase before feature weighting process. However, average accuracy results represented in Table 4 with 93.03% in the training phase and 96.48 % in the testing phase after selecting significant features using feature weight process were achieved. On the other hand, Tables 5 and 6 illustrates the prediction of email filtering using hybrid logistic regression and the two-step clustering algorithm obtained average accuracy result at 97.26% and 97.40% before feature weighting process for training and testing phases respectively. The average accuracy results after selecting significant features using feature weight process, obtained 98.41% and 98.37% for training and testing phases, respectively.

To explore the differences between this study's spam filtering technique based on the logistic regression and two-step clustering algorithms before and after improvement using weighting process and important features, an Independent Sample T-test was performed such as [29]. The achieved values can be significant if the result is below 0.05. In Table 7 the significant values are (0.006) between this study's combined LR-Two-step and LR before feature weight, and (0.0007) between the combined LR-Two-step and LR after feature weight, this indicates that the combined method reached significant enhancement on the accuracy results. Thus, a conclusion was drawn that there is a significant difference before and after feature weight and combination process. Table 7 shows the T-test statistical significance results.

Another comparison between this study's integrated technique and current approaches demonstrates in Table 8, Fig. 6 and 7. It was noted that the combined method between the logistic regression and Two-step clustering algorithm obtained best accuracy results based on both all features, and important features in the spam based dataset.

TABLE. VII. T-TEST STATISTICAL SIGNIFICANCE RESULTS

Method	**Differences between accuracy result before and after the improvement**					**T**	**Sig. Value**
	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval of the Difference**			
				Lower	Upper		
LR & LR-Two-Step (Before feature weight)	-2.444	2.150	.680	-3.982	-.906	-3.594	.006
LR & LR-Two-Step (After feature weight)	-.969	.335	.106	-1.209	-.729	-9.152	0.0007

TABLE. VIII. A COMPARISON OF THE PROPOSED METHODS AND OTHER SPAM FILTERING METHODS

Method	**Results using All Features**		**Results using Important Features**	
	Accuracy	**Error**	**Accuracy**	**Error**
Logistic Regression	**90.85%**	**9.15%**	**93.03**	**6.97**
Logistic Regression-Two-step	**93.03%**	**6.97%**	**98.41**	**1.59**
SVM [11]	90%	10%	89.34	10.66
naive Bayes [9]	78.8%	21.2%	83.17	16.83
Neural Network [13]	94.30%	5.694%	94.2	5.8
k-nearest neighbor [31]	90.8%	9.2%	88.4	11.6

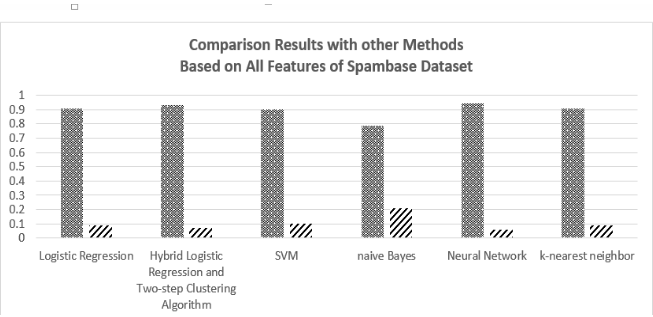


Fig. 6. A comparison between this study's proposed methods and other spam filtering methods based on all features.

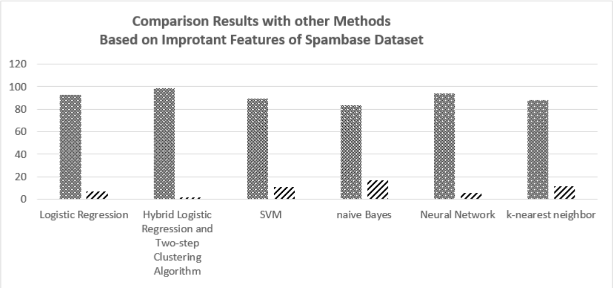


Fig. 7. A comparison of this study's proposed methods and other spam filtering methods based on important features.

Fig. 6 and 7 represent the comparison result between the proposed method and current spam classification methods. In Fig. 6, the comparison based on all spam features of spam based dataset, while Fig. 7 represents the comparison of results based on significant features that were extracted using weight feature process. It was observed that the proposed LR-two-step technique achieved the best result using both dataset features and significant features. On the other hand, the lower result was obtained by the naive Bayes method as shown in Table 8.

VI. Conclusion

This study is considered one of the main challenges through the email messages. The spammers can easily steal information by sending random spam emails via the internet. This research tried to investigate the email messages based on the logistic regression method to classify the messages to spam or non-spam. A feature weight based on the amount of data is one of the contributing parts proposed in this study to select the significant features. Another contribution is an integrated technique between the logistic regression and two-step clustering method to differentiate the email messages of spam from non-spam. The benefit of using the two-step clustering method is to group the similar emails features to study the spammers' pattern by focusing on their beavering in constructing the email messages. The proposed method used a UCI Spam base dataset to build the spam-filtering model. Based on the obtained results, conclusions were made that not all the email messages writing style features could be used by spammers. Where, only the important features that were selected using feature weight process can improve the computational time of email spam filtering. The proposed method was tested using T-test statistical significant method to prove improvement before and after feature selection and combination process. It has been shown that the LR-Two-Step can significantly enhance the filtering accuracy ratio and decrease the misfiltering error in spam dataset.

VII. Acknowledgement

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (G-421-830-37). The authors, therefore, gratefully acknowledge the technical and financial support from the DSR.

References

[1] Cormack, G.V., M.D. Smucker, and C.L. Clarke, Efficient and effective spam filtering and re-ranking for large web datasets. Information retrieval, 2011. 14(5): p. 441-465.

[2] Carpinter, J. and R. Hunt, Tightening the net: A review of current and next generation spam filtering tools. Computers & security, 2006. 25(8): p. 566-578.

[3] Schryen, G., Anti-spam legislation: An analysis of laws and their effectiveness. Information & Communications Technology Law, 2007. 16(1): p. 17-32.

[4] Ma, W., D. Tran, and D. Sharma. A novel spam email detection system based on negative selection. in Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on. 2009. IEEE.

[5] Guzella, T.S. and W.M. Caminhas, A review of machine learning approaches to spam filtering. Expert Systems with Applications, 2009. 36(7): p. 10206-10222.

[6] Massey, B., et al. Learning Spam: Simple Techniques For Freely-Available Software. in USENIX Annual Technical Conference, FREENIX Track. 2003.

[7] Zhang, L., J. Zhu, and T. Yao, An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 2004. 3(4): p. 243-269.

[8] Metsis, V., I. Androutsopoulos, and G. Paliouras. Spam filtering with naive bayes-which naive bayes? in CEAS. 2006.

[9] Marsono, M.N., M.W. El-Kharashi, and F. Gebali, Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification. Computer Networks, 2009. 53(6): p. 835-848.

[10] Marsono, M.N., M.W. El-Kharashi, and F. Gebali, Binary LNS-based naïve Bayes inference engine for spam control: noise analysis and FPGA implementation. IET Computers & Digital Techniques, 2008. 2(1): p. 56-62.

[11] Tang, Y., et al. Support vector machines and random forests modeling for spam senders behavior analysis. in IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference. 2008. IEEE.

[12] Yoo, S., et al. Mining social networks for personalized email prioritization. in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. ACM.

[13] Silva, R.M., T.A. Almeida, and A. Yamakami. Artificial neural networks for content-based web spam detection. in Proceedings on the International Conference on Artificial Intelligence (ICAI). 2012. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[14] Silva, R.M., T.A. Almeida, and A. Yamakami. Towards web spam filtering with neural-based approaches. in Ibero-American Conference on Artificial Intelligence. 2012. Springer.

[15] Largillier, T. and S. Peyronnet, Webspam demotion: Low complexity node aggregation methods. Neurocomputing, 2012. 76(1): p. 105-113.

[16] Liu, Y., et al. Identifying web spam with user behavior analysis. in Proceedings of the 4th international workshop on Adversarial information retrieval on the web. 2008. ACM.

[17] Castillo, C., et al. Know your neighbors: Web spam detection using the web topology. in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007. ACM.

[18] Rungsawang, A., A. Taweesiriwate, and B. Manaskasemsak, Spam host detection using ant colony optimization, in IT Convergence and Services. 2011, Springer. p. 13-21.

[19] Ng, A.Y., M.I. Jordan, and Y. Weiss, On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 2002. 2: p. 849-856.

[20] Aud, S., et al., The Condition of Education 2012. NCES 2012-045. National Center for Education Statistics, 2012.

[21] Chhabra, P., R. Wadhvani, and S. Shukla, Spam filtering using support vector machine. Special Issue UCCT, 2010. 1(2): p. 3.

[22] Elssied, N.O.F., O. Ibrahim, and A.H. Osman, Enhancement of spam detection mechanism based on hybrid\varvec {k}-mean clustering and support vector machine. Soft Computing, 2015. 19(11): p. 3237-3248.

[23] Elssied, N., O. Ibrahim, and A.H. Osman, A novel feature selection based on one-way anova f-test for e-mail spam classification. Research Journal of Applied Sciences, Engineering and Technology, 2014. 7(3): p. 625-638.

[24] Hopkins, M., et al., Spam base dataset. Hewlett-Packard Labs, 1999.

[25] Kaufman, L. and P.J. Rousseeuw, Finding groups in data: an introduction to cluster analysis. Vol. 344. 2009: John Wiley & Sons.

[26] Wolberg, W.H. and O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the national academy of sciences, 1990. 87(23): p. 9193-9196.

[27] Bradley, A.P., The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 1997. 30(7): p. 1145-1159.

[28] Hand, D.J., Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine learning, 2009. 77(1): p. 103-123.

[29] Osman, A.H., et al., An improved plagiarism detection scheme based on semantic role labeling. Applied Soft Computing, 2012. 12(5): p. 1493-1502.