Feature Selection Methods for Classifying Email Messages: Analysis, Proposal, and Comparative Study

Sanaa Abou Elhamayed
Department of Informatics Research, Electronics Research Institute, Cairo, Egypt
Email:sanaa-hamayed@hotmail.com
Samah Osama M. Kamel
Department of Informatics Research, Electronics Research Institute, Cairo, Egypt
Email: samah_n2003@yahoo.com

ABSTRACT

Spam Email messages have a big problem either for users or for the Internet service providers. The content of such messages may contain viruses and bad information. The spam messages also occupy a huge amount of space on the mail boxes. So, the process of Emails' classification is very important to be analyzed and discussed.

This research work aims at classifying the email messages into either spam or non-spam. The E-mail messages or a dataset can be represented in a matrix form. The rows of the matrix are representing the instances (messages) while the columns are representing the features of such instances. K-Nearest Neighbor (KNN) and Naïve Bayes (NB) are two classifiers where they are used to classify the email messages. The proposed approach based on partitioning the dataset into segment and compared with the adopted approach. Moreover, feature selection methods are adopted to choose the significant features and eliminate the others to avoid processing overheads. The choice of the relevant features plays an important role of the classification accuracy. In this work, some feature selection methods are adopted, analyzed, and operated. The performance of such methods is compared. Moreover, a feature selection method is proposed and discussed. The performance of the proposed feature selection method is compared with the adopted ones. This work is operated on a chosen dataset taken from the Internet. The dataset contains about four-thousand messages with fifty-eight features. Moreover, the dataset is supported with a target feature representing the class labels. From the practical experiments it is shown that the performance of the proposed method is better than the adopted ones. It is also expected that the proposed method is applicable to other datasets for other application domains.

Keywords - **Spam Messages, Classification Algorithms, Feature Selection Methods, Text Representation, and Performance Evaluation.**

Date of Submission: Nov 28, 2019 Date of Acceptance: Dec 24, 2019

1. INTRODUCTION

E-mail spam is undesirable message and becomes a global threat against e-mail users. E-mail spam is unrequested information sent to the e-mail boxes. Spam is considered a critical problem for users as well as the Internet Service Providers (ISPs).

Several problems are connected with e-mail spam. These include but not limited to: wasting storage space, filling users' mailboxes, consuming network bandwidth, and taking long time to delete all space messages. Spammers can use tricky methods to overcome the filtering methods like using random sender addresses and append random characters to the message subject line [4-6]. Automatic detection and/or classification algorithms are important to classify the E-mail to either spam or non-spam. The input to the classifier is the set of E-mail messages; each is described by a set of features. Feature selection approaches are imported to reduce the dimensionality of the E-mail message features. The classification phase finds the actual mapping between the training set and testing set [4-6].

The organization of this research work will be as follows: section 2 presents some of the previous work, while section 3 presents the dataset collection of e-mail messages. Section 4 presents an analysis of some feature selection methods. Section 5 discusses two classifiers mainly the k-nearest neighbor and Naïve Bayes respectively. Section 6 presents a proposed approach for feature selection based on data partitioning. Section 7outlines some measurable criteria that are used to evaluate the performance of the feature selection methods, the classification approaches, and the proposed one. Finally, section 8 presents the implementation work and discussion of results while section 9 concludes the whole work.

2. RELATED WORK

Several research efforts were done for detecting and classifying the E-mail messages. Some approaches were presented for conducting the dimensionality reduction via feature selection or feature extraction. Examples of such efforts are briefly mentioned as follows:-

[4] mentioned that the problem of high dimensionality of the feature space due to massive number of e-mails still

exists. In this concern, it is better to reduce the computational complexity and increase the classification accuracy. The authors applied a feature selection scheme based on one-way ANOVA F-test to determine the better features contributing to e-mail spam classification. That adopted scheme was used to reduce the high data dimensionality before classification. The scheme was applied and operated on a spam base dataset. The experimental work presented better results for the enhanced SVM than that of the original SVM.

[5] presented a clustering method of spam messages collected in base of anti-spam system. A genetic algorithm was developed to solve the clustering problem. The objective function is a maximization of similarity between messages in clusters which is defined by K-nearest neighbor algorithm. After classification, knowledge extraction was applied to get information about classes. Multi-document summarization method was used to get the information portrait of each cluster of spam messages. Classifying spam templates enables the authors to define the thematic dependence from geographical dependence such as what subjects prevail in spam messages sent from certain countries.

[7] mentioned that many spam detection techniques based on machine learning have been proposed. The authors proposed an optimal spam detection model based on random Forests which involves parameter optimization and feature selection. They optimized two parameters to maximize the detection rates. They provided the important features and eliminated the irrelevant ones. They decided an optimal number of selected features. They carried out their experiments on the spam base dataset which presented feasibility of their method.

[8] discussed the role of attribute selection in classification algorithms. The authors proposed an attribute selection of information gain and ranked a searching method. Each selected attribute is ranked based on the filter and entrapper method. The author used the tree-based J48 classifier with different test options namely: 10-fold cross validation, training set, supplied test set, and percentage split respectively. Labor dataset was implemented to test the adopted methods.

[9] mentioned that most classifiers based on neural networks provide results which cannot directly be interpreted as probabilities. Probabilities are important and useful for classification and misclassification. In a multiclass problem, certain misclassification results may be more expensive than others. The authors presented an extension of the extreme-learning-machine to provide probabilities as outputs for multiclass classification problems. Such information is more useful than traditional crisp classification outputs. The proposed method presented low computational time and state of the art performance.

This research work aims at classifying a collection of messages (M) or dataset into either spam messages or non-spam messages. Each message is described by a set of features or attributes. The terms: features and attributes are used interchangeably in this work. Any message m is represented using the vector space model (VSM). The VSM is used to represent any message in a vector form specifically l-dimensional vector. Assume that $M=\{m_1,m_2,\text{-----}m_n\}$ is a set of messages and $m_i=\{w_{i1}, w_{i2},\text{------}w_{il}\}$ where w_{ij} is the weight of the feature j in the message i such that ($1\leq i\leq n$ and $1\leq j\leq l$) where n and l are the number of messages and the number of features per each respectively. According to such number of features, some feature selection methods are discussed to choose the most important features. A feature selection method is proposed. The proposed and adopted feature selection methods are investigated, analyzed, and compared.

3. DATASET COLLECTION OF E-MAIL MESSAGES

To evaluate the performance of the adopted feature selection methods and classifier approaches; a dataset collection of e-mail messages is chosen and operated as a testbed. That dataset was taken from the UCI database from the center of excellence of machine intelligence in USA. Such dataset was downloaded from the website https://archive.ics.uci.edu/ml/datasets/Spambase. The dataset of the E-mail messages has about four-thousands of instances with fifty-eight attributes (features) per each. The attributes are all continuous real number ranging from 0 to 100 while the last attribute is dedicated to the class. The first 48 attributes are type of word_freq_WORD =percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the email)/ total number of words in e-mail. The attributes (49-54) are type char_freq_CHAR= percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail. While the attributes (55-57) measure the length of sequences of consecutive capital letters and the last attribute is the class label.

The class value may be 1 for the spam e-mail message and 0 for the non-spam e-mail. The non-spam e-mails are sometimes called ham messages. To do the classification process, the e-mail messages are partitioned into two parts: one for building the learning model and the other for testing. The number of instances used in building the learning model is greater than twice of those used in testing. Before handling the feature selection methods, the dataset should be screened to reject any noisy or outlier instances and this is sometimes called data cleaning.

A. Dataset Cleaning

Data cleaning is considered a preprocessing operation and can be done to reject any strange or noisy instances. The noisy instances may appear in different forms. This includes but not limited to: missing attributes, redundant instances, over-fitting and under-fitting data. The noisy attributes have a bad effect on the classification process. To make sure that the dataset is containing noisy data or not, some statistical operations can be done such as. mean, standard deviation, standard error, variance, and others. The standard division and variance are computed to reject and eliminate the noisy instances.

Variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value [https.//en.wikipedia.org/wiki/Variance]. The term Variance (var) measures how far a data set (x) is spread out.

$$var = \frac{\sum(x-\bar{x})^2}{(n-1)} \qquad (1)$$

The Standard deviation on the other hand is a measure used to quantify the amount of variation or dispersion of a set of data values [https.//en.wikipedia.org/wiki/ Standard_deviation]. Standard division (SD) means how much the members of a group differ from the mean value for the group.

$$SD = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}} \qquad (2)$$

Where n is the number of instances.

After rejecting the noisy instances from the original dataset the number of instances is reduced. Fig. 1 and Fig. 2 show the variance and standard deviation values for the original data and selected ones respectively.

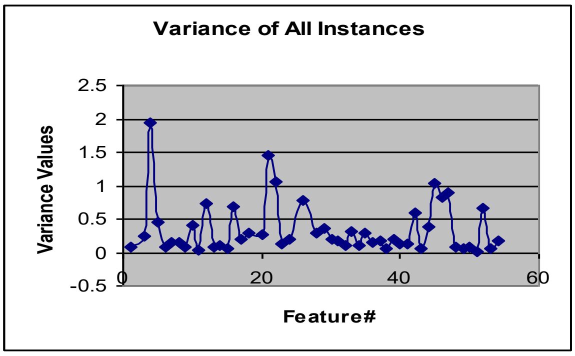

Fig.1a. Variance of All Instances

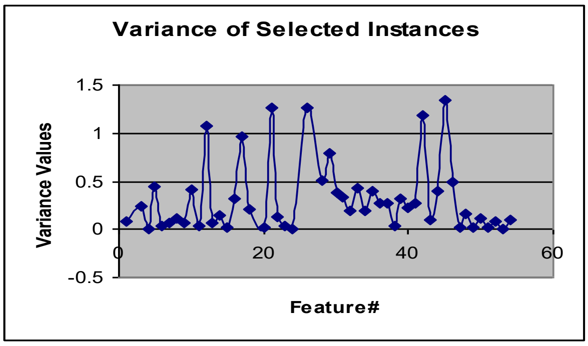

Fig.1b. Variance of Selected Instances

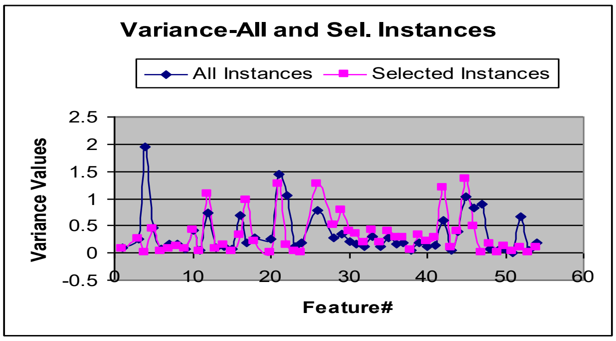

Fig.1c. Variance of All and Selected Instances

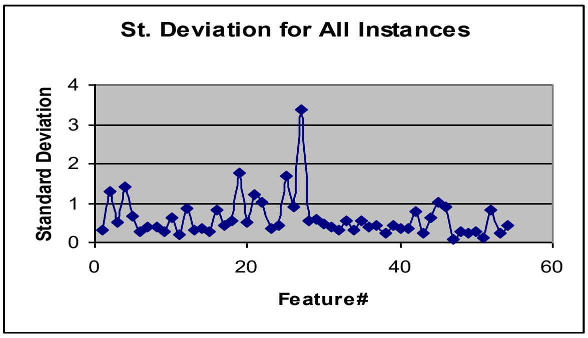

Fig.2a. Standard Deviation of All Instances

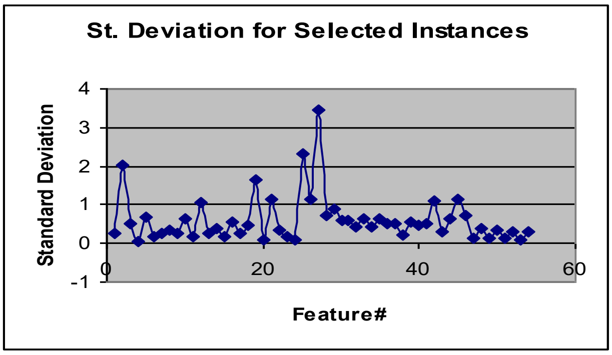

Fig.2b. St Deviation of Selected Instances

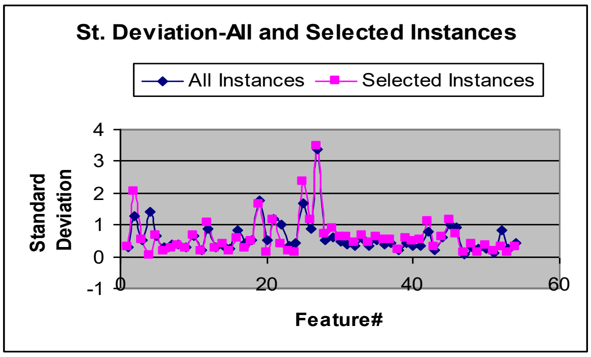

Fig.2c. St Deviation of All and Selected Instances

4. ANALYSIS OF SOME FEATURE SELECTION METHODS

The main objective of any feature selection method is to select the important and significant features (subsets) that can represent the original dataset. Feature selection methods can reduce the feature space and improve the classification accuracy.

In this work, some feature selection methods are adopted, analyzed, and operated. Such methods are based respectively on CHI-square (χ^2), correlation, feature-class similarity, and information gain. Such methods are applied as shown in the following subsections.

4.1. Feature selection based on CHI-square (χ^2)

CHI-Square (abbreviated as χ^2) is used to measure the relevance between a feature f and a class c. The high score of χ^2 means a strong relevance between f and c. i.e. the high score of χ^2 indicates that the occurrence of feature f and class c are dependent and that feature can be selected for classifying the e-mail messages. The measure between the feature f and class c can be written as follows.-

$$x^2(f,c) = \frac{t*(R_aR_d - R_cR_b)^2}{(R_a+R_c)*(R_b+R_d)*(R_a+R_b)*(R_c+R_d)} \quad (3)$$

Where t is the total number of training samples, R_a is the number of times f and c occur, R_b is the number of times f occurs without c, R_c is the number of times c occurs without f, R_d is the number of times that neither c nor f occurs.

When f and c are independent then χ^2 (f, c) =0. This means that the feature f is not containing identification information. The maximum score among all classes can be computed as.

$$x^2_{max}(f) = max_{i=1}^{c}\{(f,c_i)\} \quad (4)$$

Where c is the number of classes [10, 11].

After calculating CHI square (χ^2), different threshold values are considered to improve the classification accuracy. We have chosen five different threshold values to select the appropriate features and hence the number of chosen features for each threshold value is different from others. Fig. 3a shows the number of selected features for each threshold value. Fig.s [3b-3f] show respectively the CHI-square values w.r.t. the selected features for each threshold value.

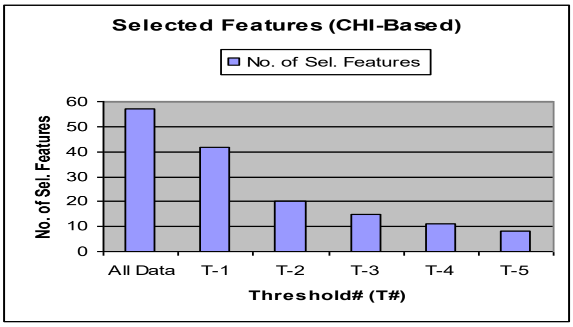

Fig.3a. Selected Features for Five Thresholds

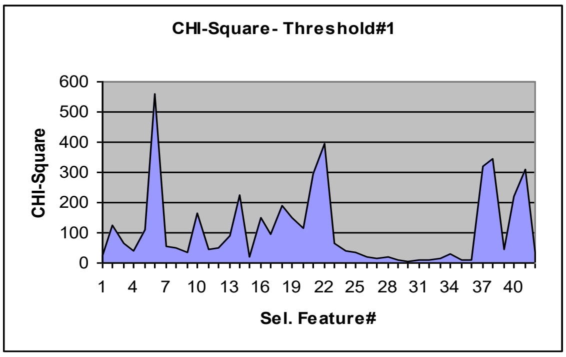

Fig.3b. CHI-Square Values for Threshold #1

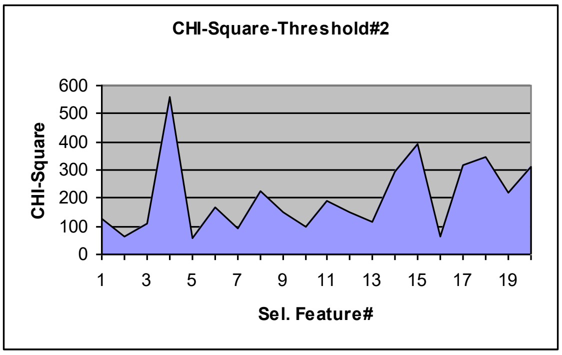

Fig.3c. CHI-Square Values for Threshold #2

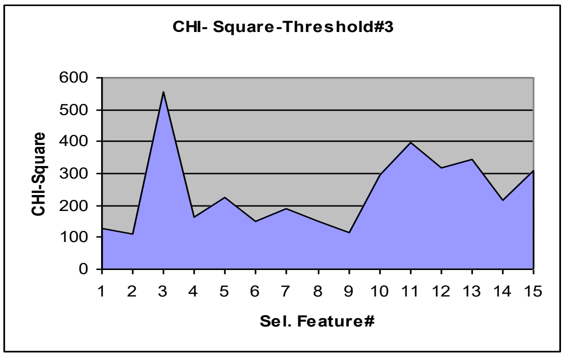

Fig.3d. CHI-Square Values for Threshold #3

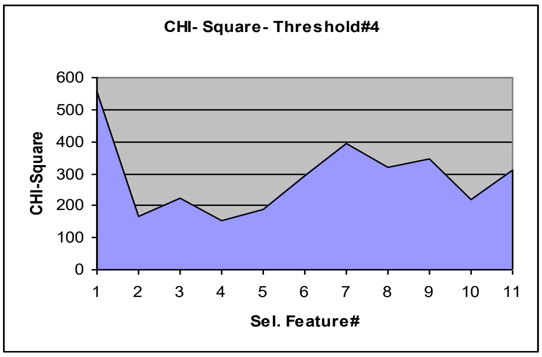

Fig.3e. CHI-Square Values for Threshold #4

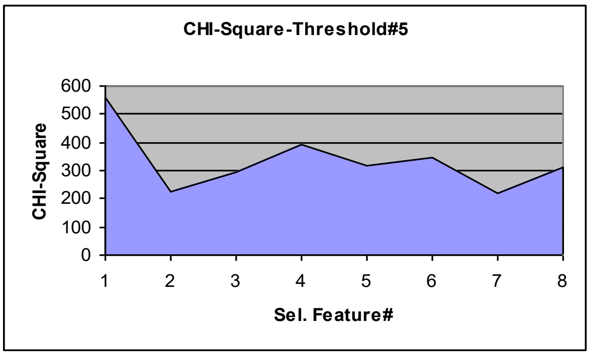

Fig.3f.CHI-Square Values for Threshold #5

4.2. Feature selection based on correlation

Feature selection based on correlation is important as it ranks a feature subset according to the correlation value. The subset of features having high correlated values with the class are extracted and taken. i.e. the features having low correlation values with the class are not significant and can be discarded as they have a bad effect on the classification accuracy [14].

Moreover, the correlation coefficient (r) can be used as an indicator for measuring the relationship between any two features or between any individual feature and the class. The correlation coefficient (r) between a feature (f) and the class c is written as shown in equation (5).-

$$r = \frac{n(\sum fc)-(\sum f)(\sum c)}{\sqrt{[n(\sum f^2)-(\sum f)^2][n(\sum c^2)-(\sum c)^2]}} \qquad (5)$$

Where. n is the number of instances, f is a feature and c is the class.

Also, five different threshold values are considered. Fig. 4a shows the number of selected features for each threshold value. Fig.s [4b-4f] show respectively the correlation value for the selected features for each threshold value.

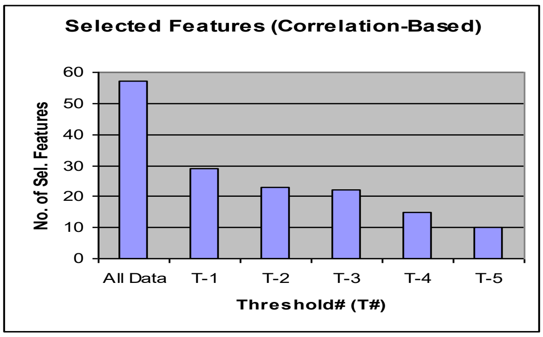

Fig.4a. Selected Features for Five Thresholds

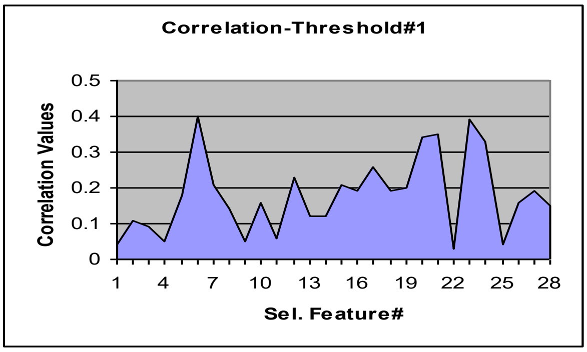

Fig.4b. Correlation Values for Threshold #1

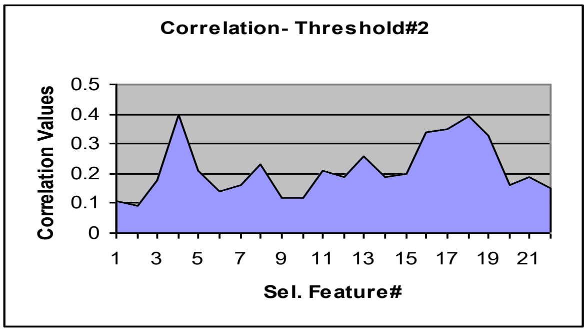

Fig.4c. Correlation Values for Threshold #2

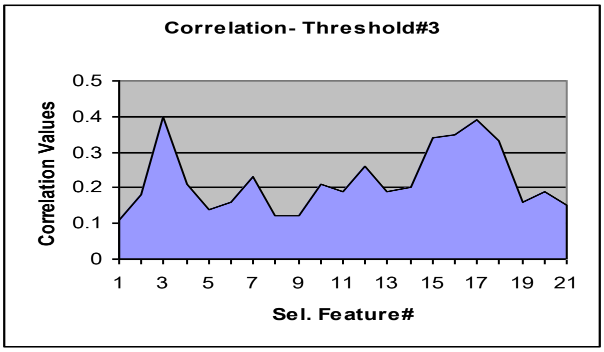

Fig.4d. Correlation Values for Threshold #3

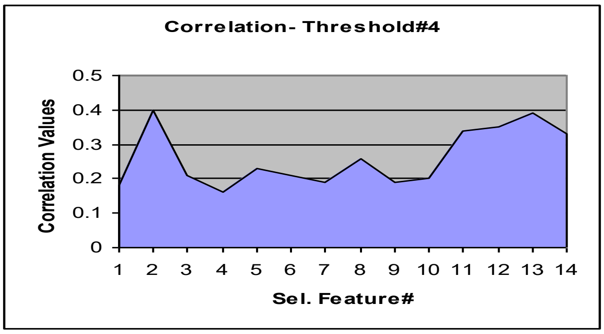

Fig.4e. Correlation Values for Threshold #4

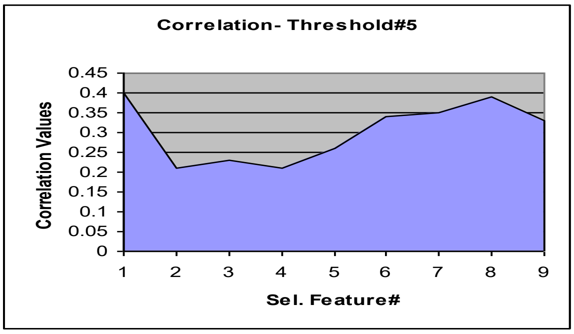

Fig.4f. Correlation Values for Threshold #5

4.3. Feature selection based on feature-class similarity

As mentioned before, an e-mail message can be represented by a set of attributes, each recording the weight of a particular feature, i.e. each message is considered an object represented in a vector form. So, the email messages of the dataset are represented as a matrix. the email messages are the rows while the features are the columns.

It is easy to find the distance and /or similarity between any feature vector and that one representing the class. The cosine angle between every feature vector and the class is the measure of such similarity. Let (for example) X and Y are respectively the feature vector and the class vector. The similarity function sim (X, Y) can be written as shown in the following formula [1].

$$sim(X,Y) = \frac{X.Y}{\|X\|.\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \qquad (6)$$

Where $\|X\|$ is the Euclidian distance norm of vector $X=(X_1, X_2, X_3,\ldots\ldots, X_n)$ where n is the number of email messages and $\|Y\|$ is the Euclidian norm of the class vector. The closer cosine value to 1 means the smaller the angle and the greater matching between feature-class vectors. Based on a threshold value (s), the features with higher similarity values above such threshold are chosen and selected while others are discarded.

Similarly as in the previous experiments, five different threshold values are considered as shown in Fig. 5.

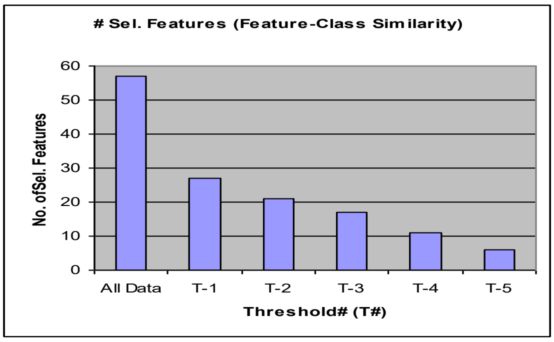

Fig.5a. Selected Features for Five Thresholds

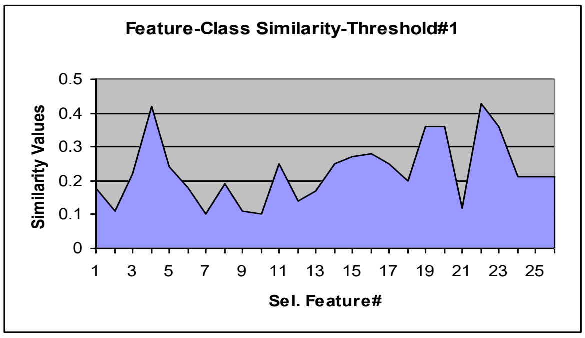

Fig.5b. Similarity Values for Threshold #1

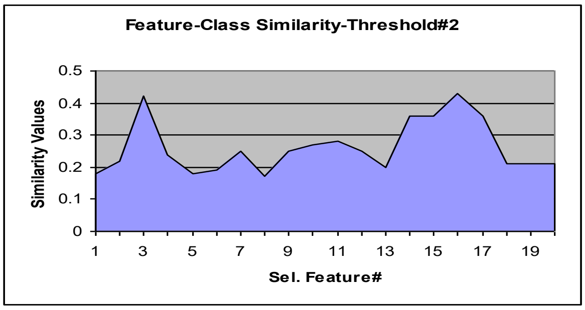

Fig.5c. Similarity Values for Threshold #2

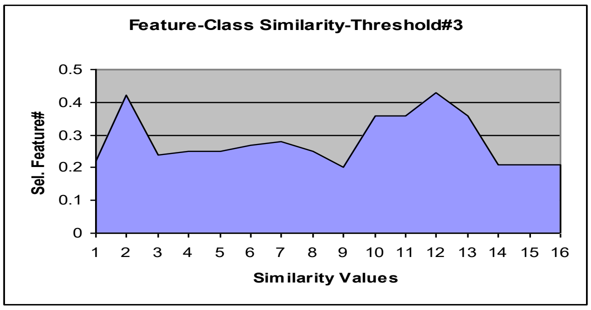

Fig.5d. Similarity Values for Threshold #3

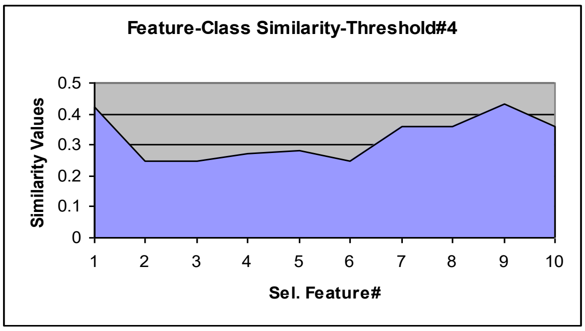

Fig.5e. Similarity Values for Threshold #4

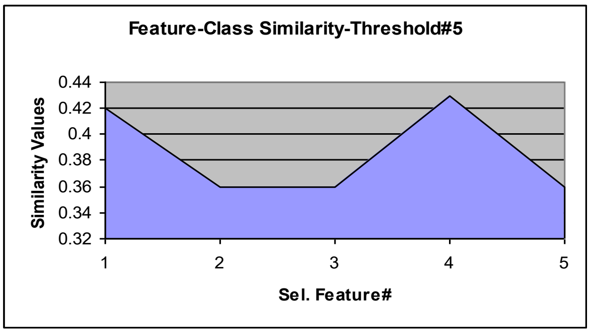

Fig.5f. Similarity Values for Threshold #5

4.4. Feature selection using information gain

As mentioned before, feature selection is done by eliminating the irrelevant and/or trivial features from the original dataset. Information gain method is adopted here to extract and select the most important features. This method measures the amount of information about the class prediction. This can take place if the only information available is the presence of a feature in the corresponding class distribution. This means that information gain (abbreviated as IG) is used to assign a maximum value to a feature if it is a good indicator for assigning the email message to any class. IG is a selection method to present score values for the features. IG for a feature f can be calculated using the following formula.

$$IG(f) = -\sum_{i=1}^{m} p(c_i)\log(c_i) + (f)\sum_{i=1}^{m} P(c_i|f)\log P(c_i|f) + P(\bar{f})\sum_{i=1}^{m} P(c_i|\bar{f})\log p(c_i|\bar{f}) \quad (7)$$

Where m is the number of classes, P (c_i) is the probability of the class c_i, P (f) and p $(\bar{f})$ are the probabilities of presence and absence of the feature f respectively. P and p are respectively the conditional probabilities of class c_i given the presence and absence of feature f [1-3].

As mentioned in the previous experiments, five threshold values were chosen and the results are shown in Fig. 6.

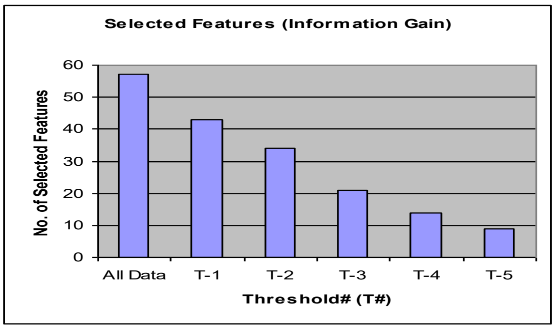

Fig.6a. Selected Features for Five Thresholds

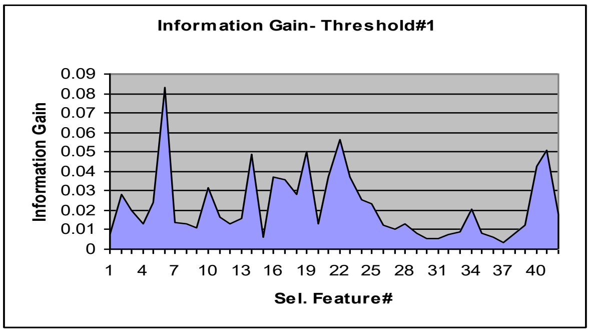

Fig.6b. Information Gain for Threshold #1

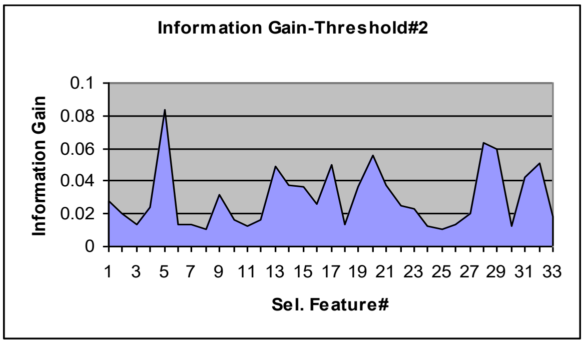

Fig.6c. Information Gain for Threshold #2

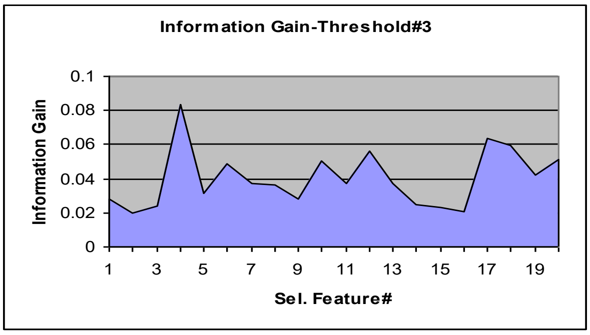

Fig.6d. Information Gain for Threshold #3

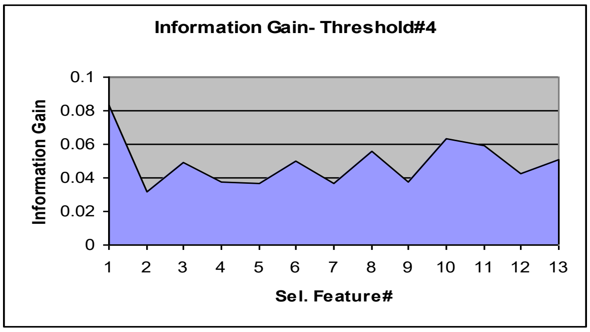

Fig.6e. Information Gain for Threshold #4

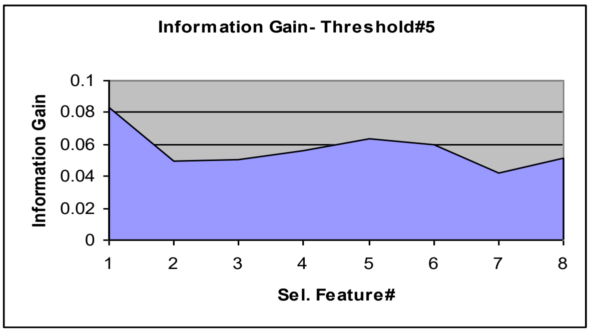

Fig.6f. Information Gain for Threshold #5

5. CLASSIFICATION OF E-MAIL MESSAGES

In this section, two classifiers are adopted, analyzed, and operated on the dataset. Such classifiers are used to classify an input instance to predict it is spam or non-spam emails. The adopted classifiers are. K-Nearest Neighbors (KNN), and Naive Bayes (NB). Such classifiers are briefly presented in the following subsections.

5.1. Classification of E-mail messages using the K-nearest neighbor

The K-nearest neighbor (KNN) is a common approach for determining a class to which an object belongs to. The objects in this work are either spam e-mail messages or ham e-mail messages. The inputs to the classifier are those selected features produced by the previous selections methods. The output from the classifier is the class name of an input e-mail message. Spam or non-spam. The KNN classifier is used with Euclidean Distance and 10 nearest neighbor. Also, 10-fold cross-validation is used. The Euclidean distance between each feature and the class is calculated by the following equation.

$$d(f,c) = \sqrt{\sum_{i=1}^{n}(f_i - c_i)^2} \qquad (8)$$

Where f is a feature, c is a class, and i is the number of instances.

5.2. Classification of E-mail messages using naïve bayes classifier

Naïve Bayes classifier is one of the common classifiers used in classifying several datasets in different applications. It is considered as one of the statistical classifiers which can predict class membership probabilities where the probability of a given instance falls into a specific class. The Naïve Bayes classifier is based on Bayes' Theorem [23].

The Naïve Bayes (abbreviated as NB) is powerful, easy, and a language independent approach. To classify a dataset of email messages, the following equation is adopted [24].

$$p(class|message) = \frac{p\ (message)p(message|class)}{p(message)} \qquad (9)$$

Where P(class | message) is the probability of class given a message, or the probability that a given message m belongs to a given class c. P(message) is the probability of a message. P(class) is the probability of a class where it can be computed from the number of messages in the category divided by the message number in all categories. P(message | class) represents the probability of a message given the class. Moreover, any email message can be represented by a set of words, where P(message | class) can be written as

$$\text{P(message|class)}=\prod_i P(word_i \mid class),$$

,so (10) P(class|message)=P(class) $\prod_i P(word_i \mid class)$ (11)

Where P(word$_i$ | class) is the probability that the i-th word of a given message occurs in a message from the class c. For more details, the reader can refer to [24].

6. A PROPOSED APPROACH FOR FEATURE SELECTION

As mentioned before, feature selection is used in several applications and/or tasks where the machine learning is one of them. Usually, the huge dataset contains a large number of features and contains at the same time some irrelevant or redundant information. To improve the performance of any classifier, such redundant and irrelevant information should be avoided. i.e. the existing redundant and irrelevant information/ features can downgrade the learning accuracy and deteriorate the performance of the learning models. So, feature selection is an important issue in the adopted domain of machine learning. This is because feature selection can directly affect the classifier's accuracy and generalization performance. As mentioned before, feature selection aims at finding a feature subset that has the most discriminative information from the original feature set.

Moreover, there are different types of feature selection methods; the filtering method is that one adopted in this work. The filter-based feature selection method aims to select the best feature subset based on the intrinsic properties of the data. [18-20].

The data layout of the proposed approach can be represented as a collection of vectors or a matrix X. The dataset can be written as [X: x_{ij} where $1 \leq i \leq m$, and $1 \leq j \leq n$] where m is the number of instances of the dataset each with n features. Moreover, the universal dataset features can be imagined as the set U of possible features and it is required to identify the subset of features $F \subset U$ where the features in F have the capability to build a model to best predict the target class [19]. The proposed approach for selecting features is based on data partitioning as shown in the following subsections.

6.1. Feature selection based on data partitioning using the correlation measure

This method is based on partitioning the dataset into a number of equal size segments N_s where N_s is the value of total number of dataset instances (m) divided by the segment size (s), where m and s are the number of instances for the dataset and segment size respectively. Each segment is run and operated individually to find a dependency measure between X values and the response variable of the class label. The dependency measure between any feature in X and the class label Y in our case is the correlation. That correlation (r) is considered a statistical measure. It is important to mention that r(X,Y)=0 if and only if X and Y are independent. The formula of correlation is that one mentioned in equation (5). This means that n correlation values are computed individually for the n features and the class label. This is done for each segment or data partition.

The same process is done N_s times where N_s is the number of segments.

From the above steps N_s values for the correlation of each feature with the class label are obtained. Let the correlation values are r_{ij} for $1 \leq i \leq N_s$ and $1 \leq j \leq n$.

The average of the correlation values ($\bar{r}_i$)for each feature is computed.

$$\text{Average of } \bar{r}_i = \sum_{j=1}^{n} \frac{r_{ij}}{N_S} \text{ where } 1 \leq i \leq N_s \qquad (12)$$

A threshold value is defined. The features with average correlation values above that threshold are chosen. This means that different local maximum of the average correlation between those features and the target class label are selected.

The threshold value in the previous step is changed and every time the classification accuracy is calculated and reported.

The final selected features are those that present the maximum accuracy values.

This approach was applied and tested on the same dataset as shown in the next section.

7. MEASURABLE CRITERIA FOR PERFORMANCE EVALUATION

To evaluate the performance of the conducted feature selection methods, the adopted classifier approaches, and the proposed one some criteria are considered. The criteria are accuracy, precision, recall, and F-measure [17].

$$accuracy = \frac{number\ of\ e{-}mails\ correctly\ categorized}{total\ number\ of\ e{-}mails}$$

(13)

$$precision = \frac{TP}{TP+FP} \qquad (14)$$

$$recall = \frac{TP}{TP+FN} \qquad (15)$$

$$F - measure = \frac{2*(precision*recall)}{precision+recall} \qquad (16)$$

Where, TP is the true positive which represents the number of correct classifications or predictions that an instance is positive. FP is the false positive which represents the number of incorrect classifications/predictions that an instance is positive. FN is the false negative which is the number of incorrect predictions that an instance is positive. TN is the true negative which is the number of incorrect predictions that an instance is negative.

7. IMLEMENTATION WORK AND DISCUSSION OF RESULTS

This work presents an analysis and investigation of some feature selection methods. Before discussing the performance of the adopted methods, let us briefly mention some sort of screening of the adopted dataset. The dataset contains thousands of instances of email messages. It is important to reject any outlier instances that may exist in that dataset. The outlier instances may be those instances with over-fitting or under-fitting values. The outlier instances may be also those instances that are redundant or those ones with some missing data. The outlier instances are very important to be rejected to overcome their bad impact on the performance of the feature selection methods. So, two statistical operations mainly the standard deviation and variance were consulted to detect any noisy or outlier instances. Such statistical operations are important to reject the abnormal instances.

The adopted feature selection methods are based respectively on CHI-Square, correlation, feature-class similarity, and information gain. Each method was operated and tested on the dataset before and after rejecting the outlier instances. It is shown that the performance of all methods using the cleaned data outperform those corresponding values of the data with noisy instances. The performance of each method changes by changing the number of selected features due to changing the threshold values. From the practical work, the performance of the feature selection based on correlation is the best. This is clear for the two adopted classifiers namely: K-Nearest neighbor and Naïve Bayes.

Moreover, the KNN and NB classifiers presented before are operated on the adopted dataset. The classifiers are also operated on the dataset after cleaning the noisy instances. Noisy instances include those over fitting instances, under fitting instances, and missing attributes instances. The precision, recall, F-measure, and accuracy of the two classifiers are shown respectively in Fig.s 7 for the original dataset and the cleaned one. It is shown that the performance of NB outperforms that of the KNN classifier. This is clear in the evaluation criteria specially the accuracy. This is also valid for either the original data or the cleaned one. Moreover, the performance of feature selection methods based on CHI-Square, correlation, feature-class similarity, and information gain were also reported. The evaluation criteria for such feature selection methods are shown respectively in Fig.s 8a, 8b, 8c, and 8d using the two classifiers. It is shown from Fig. 8 that the precision, recall, F-measure, and accuracy are better for the NB classifier than those of the KNN. For each feature selection method, five experiments were operated using different threshold values. The number of selected features is changed after changing the threshold value. The performance of the adopted evaluation criteria is also changed by changing the threshold value. This is valid for all the adopted feature selection methods. Moreover, the performance of the feature selection method based on correlation has the best performance compared to the other presented methods.

The proposed approach considered different threshold values. In our experiments, five threshold values were taken. The number of selected features is changed as shown in Fig. 9a when the threshold value is changed. Moreover, the Naïve Bayes classifier was operated on the selected features for each threshold value. As a result, the classification accuracy is changed by changing the number of selected features. This is shown in Fig. 9b.

Moreover, the proposed approach is based on partitioning the dataset into a set of segments. Each segment is operated as a testbed using the correlation value between the class and every individual feature. The accuracy of the proposed method is reported for each experiment where the threshold value is changing. From the obtained results, it is clear that the number of selected features changes by changing the threshold value. As a result, the classification accuracy changes by changing the number of selected features. Moreover, the accuracy values for the proposed method outperform those other adopted ones. The peak value of accuracy for the proposed method is better and higher than those peak values of the adopted methods as shown in Fig. 10. The accuracy values are computed for the adopted feature selection methods and the proposed one using the Naïve Bayes classifier. This is because the performance of Naïve Bayes classifier is slightly better than the corresponding values of the K-Nearest Neighbor classifier.

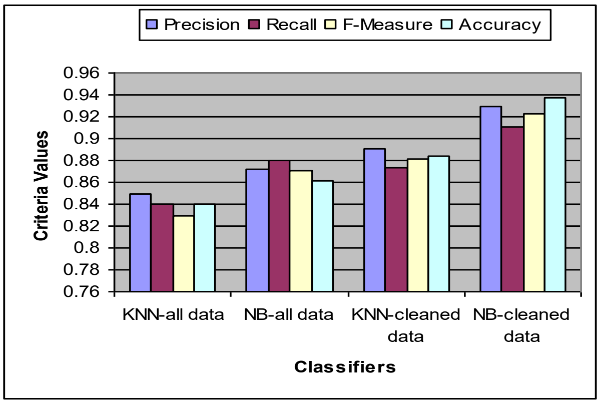

Fig.7. Classifiers Operated on All and Cleaned Data

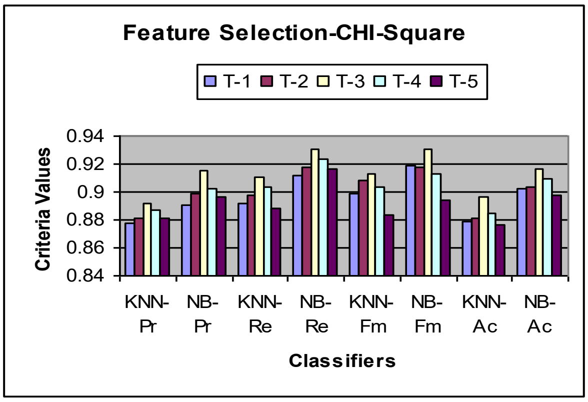

Fig.8a. Criteria Values Using CHI-Square

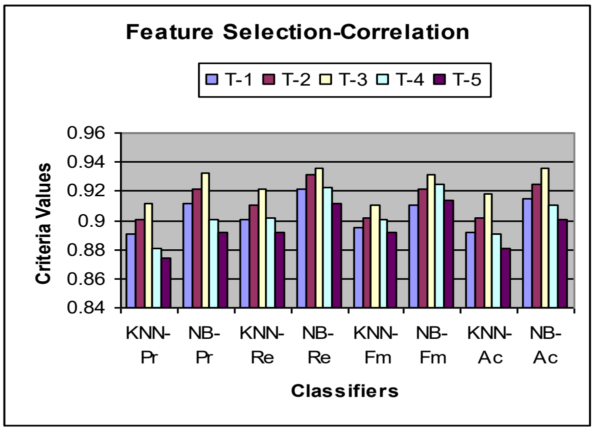

Fig.8b. Criteria Values Using Correlation

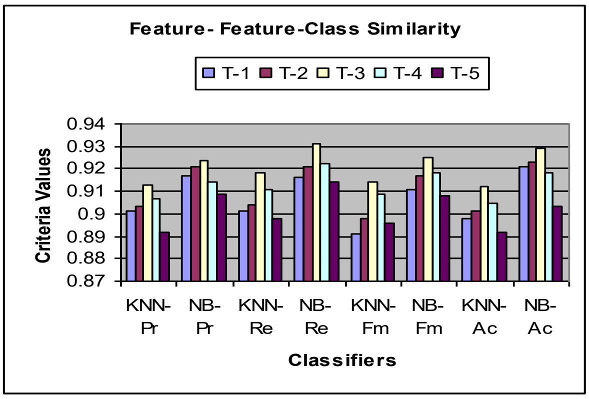

Fig.8c. Criteria Values Using Feature-Class Similarity

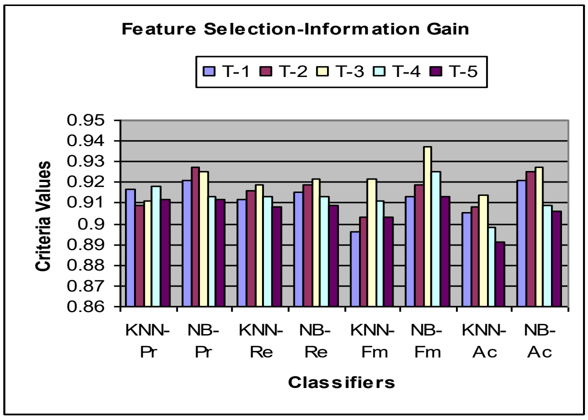

Fig.8d. Criteria Values Using Information Gain

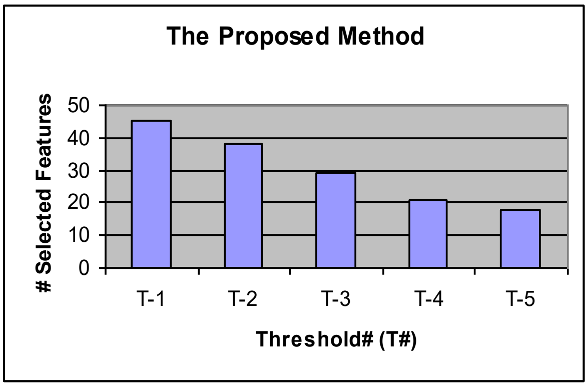

Fig.9a. The Proposed Feature Selection Method

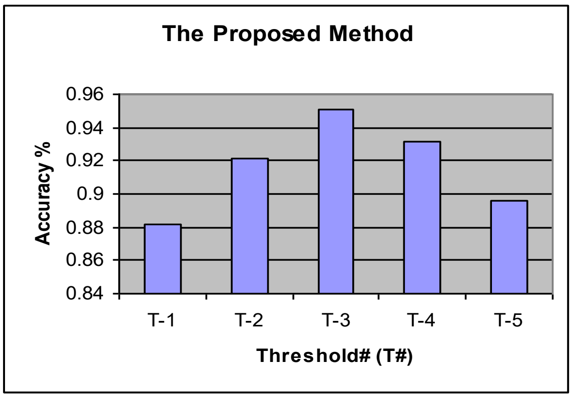

Fig.9b. Accuracy% for Five Threshold Values

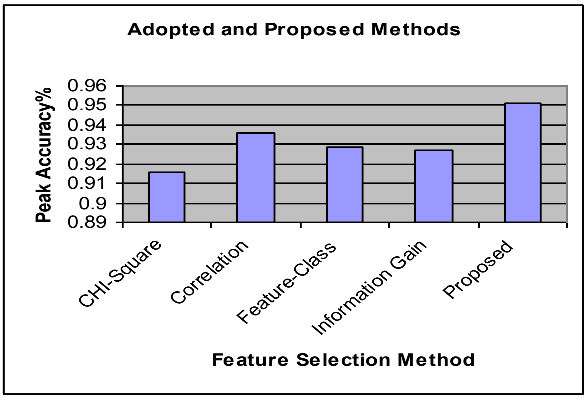

Fig.10. Peak Accuracy% for the Adopted Methods

8. CONCLUSION

This work analyzed and investigated some feature selection methods for classifying the email messages into either spam or non-spam emails. The methods were operated and tested using a dataset of thousands of email messages. Moreover, a proposed approach based on partitioning the dataset into a set of segments was also operated and tested. A comparative study among the performance of the adopted methods and the proposed one was presented. The accuracy of the proposed method outperforms the other adopted approaches. All the adopted methods as well as the proposed one presented better performance after rejecting any outlier instances than the corresponding values for the dataset without cleaning. The threshold values play an important role in the accuracy values. The accuracy value changes by changing the number of selected features. The accuracy values were decreased when using all the dataset features and also when the number of selected features is low.

Moreover, the maximum accuracy value for the proposed approach was higher and better than those maximum values of the other adopted methods.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques: the 3rd Edition", 2013.

[2] Alper K U, An Improved Global Feature Selection Scheme for Text Classification, *Expert Systems with Applications, 43*, 2016, 82-92.

[3] Pradnya K and Manisha M, A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification, the International Journal of Science and Research, 5(5), 2016, 1267-1275.

[4] Nadir E, Othman I, and Ahmed O, A Novel Feature Selection Based on-Way ANOVA F-Test for E-mail Spam Classification, The Research Journal of Applied Sciences, Engineering and Technology, 7(3), 2014, 625-638.

[5] Rasim M. A, Ramiz M. A, and Saadat A. N, Classification of Textual E-mail Spam using Data Mining Techniques, the Applied Computational Intelligence and Soft Computing, 2011, Article ID416308, 8pages, Hindawi Publishing Corporation, 2011.

[6] W. A. Awad and S. M. Elseuofi, Machine Learning Methods for Spam E-mail Classification, the International Journal of Computer Science and Information Technology, 3(1), 2011, 173-184.

[7] Samy M L, Dong S K, Ji Ho K, and Jong S P, Spam Detection using Feature Selection and Performance Optimization, The International Conference on Complex, Intelligent and Software Intensive Systems, Sponsored by IEEE Computer Socity, 2010, 883-888.

[8] S. Dinakaran and P. Thangaiah, Role of Attribute Selection in Classification algorithms, The International Journal of Scientific and Engineering Research, 4(6, 2013, 67-71.

[9] Emil E, Andrey G, Anton A, Kaj Mikel, Yoam M, Dusan S, Rui N, Bo He, and Amaury Lendasse, Exterme Learning Machine for Multiclass Classification: Refining Predictions with Gaussian Mixture Models, Downloaded from the Internet in 2016 from the Website http//:pdfs. semanticscholar.org/be7b/de5ae0ad2c2c9773cc7a0211e9fda241e955.pdf

[10] B. S. Harish and M. B. Revanasiddappa, A Comperhensive Survey on Various Feature Selection Methods to Categorize Text Documents, The International of Computer Application, 164(8), 2017, 1-7.

[11] Songtao S, Minyong S, Wenqian S, and Zhiguo Hong, Improved Feature Weight Algorithm and its Application to Text Classification, Hindawi Publishig Corporation: Mathematical Problems in Engineering, Vol. 2016, Article ID 7819626, 1-12, http"//dx.doi.org/10.1155/2016/781962.

[12] Wenyan Z, Xuewen L, and Jingjim W, Feature Selection for Cancer Classification using Microarray Gene Expression Data, Biostatistics and Biometries:Open Access Journal, 1(2), 2017, 1-7.

[13] Noga L and Lior W, Minimal Correletaion Classification, Downloaded from the Internet in 2017 from the website http:// www.cs.tau.ac.il/~wolf/papers/minCorr.pdf

[14] Veronica B C, Noelia S M, and Amparo A B, A Review of Feature Selection Methods on Synthetic Data, The Knowledge Information System, 2013, 34:483-519, DOI 10.1007/s10115-012-0487-8.

[15] M.Vasantha and V.Subbiah Bharathy, Evaluation of Attribute Selection Methods with Tree based Supervised Classification-A Case Study with Mammogram Images, International Journal of Computer Applications (0975 – 8887), 8(12), 2011, 35-38.

[16] Mark A. H and Geoffrey H, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, IEEE Transactions on Knowledge and Data Engineering, 15(3), 2003, 1-16.

[17] Muhammad Z A, Aurangzeb K, Shakeel A, Maria Q, and Imran A K, Lexicon-Enhanced Sentiment Analysis Framework using Rule-Based Classification, the Journal Plos ONE, , 2017, 1-22.

[18] Jianzhong W, Shuang Z, Yugen Yi, and Jun K, An Improved Feature Selection Based on Effective Range for Classification, The Scientific World Journal, Hindawi Publishing Corporation, Article ID-972125, 2014, 1-8.

[19] Yingbo Z, Utkarsh P, Ce Zhang, Hung N, Xuan L N, Re Christopher, and Venu G, Parallel Feature Selection Inspired by Group Testing, Downloaded From the Internet in 2017 From the Website https://cse.buffalo.edu/~hungngo/papers/nips14.pd

[20] Claudio R, Yann-Ael Le B, and Gianluca B, Feature Selection in High-Dimensional Dataset Using MapReduce, PP. 1-9, Downloaded From the Internet in 2018 From the Website https://link.springer.com/chapter/10.1007/978-3-319-76892-2_8.

[21] M.F. Zaiyadi, and B. Baharudin, Proposed Hybrid Approach for Feature Selection in Text Document Categorization, The World Academy of Science Engineering and Technology, International Journal of Computer and Information Engineering, 4(12), 2010, 1799-1803.

[22] Jose L. Torrecilla and Alberto S, Feature Selection in Functional Data Classification with Recursive Maxima Hunting, The 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016, 1-9.

[23] Subrata K M, "erformance Analysis of Data Mining Algorithms for Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree, The International Journal of Engineering and Computer Science, 6(2), 2017, 20388-20391.

[24] Fadi T, Mohammad A E, Mannam Z, and Wael M H, Naïve Bayesian Based on Chi-Square to Categorize Arabic Data, The Communications of the IBIMA, 10, 2009, 158-163.

[25] Ashok Badresiya, Saifee Vohra, and Jay Teraiya, Performance Analysis of Supervised Techniques for Review Spam Detection, International Journal of Advanced Networking Applications (IJANA), 2014, 21-24.

Biographies and Photographs:

Sanaa Hassan Abou Elhamayed is a PhD of engineering holder from Cairo, Egypt. She works as a part of Informatics Research Department in ERI. Her research interests are natural language processing, nformation system, and machine learning. She is a researcher in Electronic Research Institute. List of her latest publication: Enhancement of agriculture classification by using different classification systems. International Journal of Computer Applications (IJCA), 2016. Classifying datasets using some different classification methods. International Journal of Engineering and Technical Research (IJETR), 2016. Comparative Study on Different Classification Techniques for Spam Dataset. *International Journal of Computer and Communication Engineering* (*IJCCE*), 2016.

Samah Osama M. Kamel is Researcher in Dep. of Informatics at Electronic Research Institute, Egypt. Received B.S. degree in electronics and communications from Zagazig Faculty of Engineering, Zagazig University, in 2001. M.SC. “Secure IP Telephony Attack Sensor" at the Faculty of Engineering – Ph.D “Wireless Network Security System Analysis and Design” at the Faculty of Engineering–Shoubra, Electrical and Communication Engineering Department. There are many researches in network security and information security, machine learning and deep learning.