Email classification analysis using machine learning techniques

Khalid Iqbal and Muhammad Shehrayar Khan
Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock, Pakistan

Received 18 January 2022
Revised 15 March 2022
20 March 2022
15 April 2022
Accepted 17 April 2022

Abstract

Purpose – In this digital era, email is the most pervasive form of communication between people. Many users become a victim of spam emails and their data have been exposed.

Design/methodology/approach – Researchers contribute to solving this problem by a focus on advanced machine learning algorithms and improved models for detecting spam emails but there is still a gap in features. To achieve good results, features also play an important role. To evaluate the performance of applied classifiers, 10-fold cross-validation is used.

Findings – The results approve that the spam emails are correctly classified with the accuracy of 98.00% for the Support Vector Machine and 98.06% for the Artificial Neural Network as compared to other applied machine learning classifiers.

Originality/value – In this paper, Point-Biserial correlation is applied to each feature concerning the class label of the University of California Irvine (UCI) spambase email dataset to select the best features. Extensive experiments are conducted on selected features by training the different classifiers.

Keywords Neural network, Machine learning classifiers, Feature selection, Spam email, UCI

Paper type Research paper

1. Introduction

There are many tools for communication on the Internet. One tool that is used to convey your message to others more formally is called email. Spam is a very complicated problem in email services. Spam email is an unwanted and unwelcome email sent to users which contains job offers, selling products, services and so forth. More than 85% of spam emails are sent to users [1]. Email is not only used for personal communication but also for resolving queries of clients, job handling tasks and social activities. The email categorization as spam or not spam is mostly based on the body of the email in machine learning. The specific keywords used in the body can identify the spam email. As for the detection of spam email or text, different types of model and feature selection methods are used like structural and social network features [2], genetic search algorithm for feature selection [3] and Infinite latent feature selection [4].

In this paper, the primary focus is on the selection of features to get better performance for the classification of spam and ham emails. The major contributions in this research are as follows:

(1) We experimented with Distance-based (K-Nearest Neighbor (KNN), Support Vector Machine (SVM)), Tree-based (Random Forest (RF), Decision Tree (DT)) and Gradient-based (Artificial Neural Network (ANN), Logistic Regression (LR), Radial Basis Function (RBF)) algorithms on the University of California Irvine (UCI) spambase email dataset.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-01-2022-0012

(2) We select the Point-Biserial feature selection technique to extract the most relevant features for the classification purpose of spam email.

(3) To the best of our knowledge, no research study shows that this feature selection technique is used for spam email classification.

(4) We experiment on extracted relevant features with the Distance-based, Gradient-based and Tree-based algorithms.

Point-Biserial correlation is used to measure the relationship between the class labels with each feature. We use the dataset in which features are continuous and class labels are nominal in 1 and 0. The Point-Biserial correlation is used to measure the relationship between a continuous variable and binary variable that supported and suited the dataset we used in this research. The ANN is applied to UCI spambase email dataset to get the best result, but the problem is that the feature selection technique is not used to select the best features from data for the applied model [5]. If the algorithm is applied to data without preprocessing, it leads to less accuracy. The feature selection and dataset size are the factors that contribute to the accuracy of the machine learning model. The Support Vector Machine is applied to a dataset having 400 emails for training, if the lesser the data then the greater is the chance of the overfitting. The training accuracy of the data may be high and lower accuracy of testing data is achieved [6]. Overfitting occurs when the complex model is made for a simple dataset. Mostly spam email affect the user in the form of time consumption while reading spam email, bandwidth and in form of space that is required for the storage of spam email [7]. Users spend a lot of time reading spam emails which are useless for them. Due to the above reasons, this problem is considered in this article to find a better solution to it. Already some articles were published on this issue by using different techniques, but the maximum accuracy achieved by these articles is 95% due to the abovementioned reasons. The machine learning algorithms can solve these issues by taking minimum time [8].

To handle these issues mentioned above in the proposed methodology, a large dataset is taken consisting of about 5000 instances and try to minimize the chances of overfitting. For the improvement of the accuracy model, feature selection techniques are applied to preprocess the data before applying the machine learning model. For the evaluation of the model, 10-fold technique is applied and the overfitting and underfitting of the applied model are checked.

The remainder of the paper is organized as follows: Section 2 is the literature review that describes the details about the previous paper on the selected topic. Section 3 represents the proposed methodology for the classification of spam emails. Section 4 is the experiment and results which include the evaluation of all the techniques used in classification. Section 5 is the conclusion section which describes the conclusion of this paper and the achievement of experiments performed.

2. Literature review

The spambase UCI dataset was used for the classification of spam and ham emails. The Infinite latent feature selection was used for the selection of features. The authors were applied ten machine learning algorithms for the performance comparison between them which were RF, ANN, Logistic, SVM, RF, KNN, DT, Bayes Net, NB and RBF [4]. The strength of the suggested work is as follows:

(1) The author used Distance-based, Gradient-based and Tree-based machine learning algorithms.

(2) To reduce the influence, biasness of features having large value normalization is performed.

(3) For splitting, 10-fold cross-validation is used.

The weakness is that authors did not mention the number of neurons used in ANN; while using less number of neuron for complex problem in ANN cause to give less accuracy as compared to other algorithms. ANN could learn by itself; this quality of ANN is not present in Distance-based algorithms. For complex models, increasing the number of neurons in ANN improves the performance of classification.

The spam Assassin dataset was used and applied 24 different machine learning classifiers by using the Weka tool and achieved an accuracy of 96.32% which is the highest accuracy among other classifiers. The strength of Sharma and Amit's work is they used a large variety of machine learning algorithms to measure their performance individually. The weakness in the proposed methodology [9] is not using any feature selection technique to select the more relevant features among others features. Six hundred mails were used in the in the filtration of the spam emails. Of which, 400 mails were used for testing data and the remaining 200 were used for training data. The weighted Support Vector Machine classifier got 99.5% accuracy. The weakness is the dataset includes only a smaller number of emails as compared to spambase UCI's thousands of emails for experimentation. Less number of instances in any dataset have higher chances of accurate results. The researchers must have enough data to test well the model of machine learning algorithms.

Details of different papers are given in Table 1 according to the dataset they used and which algorithms perform best on these datasets for email classification. The Enron dataset is downloaded for the classification of spam emails and the classifier implemented here were J48 and Multilayer Perceptron which belong to the artificial-neural network family. J48 and Multilayer Perceptron achieved an accuracy of 93% and 92%, respectively [11]. The experimentation is performed with data of two different sizes, one was 1000 mails size and the other was 5000 mails size. Three classifiers were implemented which were Support Vector Machine, Naive Bayes and J48. When 1000 mails size was used, SVM, NB and J48 achieved 92, 97.2 and 95.8% accuracy, respectively [12]. When 5000 mails size was used, Support Vector Machine and Naive Bayes accuracy dropped by 1.8% and 0.7%, respectively. J48 was increased by 1.8%. The spambase UCI dataset was used for the classification of spam emails. Five different experiments were performed and 96.4% accuracy was achieved using EDT [6].

Table 1. Review of machine learning classifiers

References	Classifiers	Result
[4]	Random Forest, Support Vector Machine, Artificial Neural Network, Naive Bayes, Radial Basis Function, Bayes Net, Logistic Regression, Decision Tree	95.45%
[6]	Decision Tree, Support Vector Machine, Neural Network, Naive Bayes, Euclidean distance transformation (EDT)	96.4%
[9]	Bayes Net, Random Forest, J48, Multilayer Perceptron, Random Committee, Random Tree, Kstar	96.23%
[10]	Support Vector Machine	99.5%
[11]	J48 and Multilayer Perceptron	93%
[12]	Support Vector Machine Naive Bayes and J48	95.8%
[13]	RIPPER	95%
[14]	J48, IBK and Naive Bayes	96.3%
[15]	Artificial Neural Network	85.31%
[16]	Naive Bayes	89.7%

The spam and ham emails were recognized using two different emails sizes 400 and 50 [13]. Repeated incremental pruning to produce error (RIPPER) reduction technique was used to classify emails. When 400 mails size was used, 90% accuracy was achieved using RIPPER and when 50 mails size was used, 95% accuracy was achieved on the email dataset. The Facebook dataset was used for the identification of spam messages. J48, IBK and NB classifiers were applied to the Facebook dataset and as compared to these three classifiers J48 produced good results [14]. The spambase UCI dataset was used for the classification of ham and spam emails, features were selected from the spambase UCI dataset using the feature selection technique which is called Infinite latent selection [4]. Ten machine learning classifiers were implemented here, and the results showed that the RF classifier achieved the best accuracy as compared to others. Accuracy of 95.45% was achieved using the RF technique. Maximum accuracy was achieved using the spambase UCI dataset and pass-through Multilayer Perceptron with sigmoid function [15]. This method proved that it correctly classifies the email spam more than 85.31%. A total of 4601 email records were used in it, of which 3233 emails were used for training the model and the remaining 1368 emails were used for testing the model which was 30% of all the datasets. The spam messages were used for the filtration of spam emails [16]. Five different versions of NB were implemented on fresh spam messages. Accuracy was achieved by implementing a two-stage smoothing version that is highest than the others. Based on the previous articles published on the classification of spam emails, we conclude some points as follows:

(1) Many wide and effective classifiers for ham and spam classification of emails have been introduced.

(2) Different articles use different types of datasets related to spam email.

(3) Mostly 30 and 20% of whole dataset instances are used for testing and the remaining dataset instances are used to train the specific model.

(4) All the datasets have some gaps which can be fulfilled by preprocessing the data and as well as different feature selection techniques are used for ham and spam email classification.

(5) Till today, many researchers are working on different datasets and using different classifiers to improve the results and achieve the best accuracy of all the others.

(6) Ham and spam email classification is still an open research question.

(7) The greatest accuracy is achieved using spambase UCI datasets, 95.45%.

Some issues of the literature discussed above are overcome in the proposed methodology and the main points are as follows:

(1) The latest and larger dataset of spam and ham emails than the existing ones is used for experimenting with the proposed methodology because the actual evaluation of any proposed methodology can measure on a larger dataset.

(2) The small dataset has a higher probability of achieving good accuracy using even simple methods but it is hard to achieve better accuracy in the larger dataset.

(3) The Point-Biserial feature selection technique is used to find the features that have a relation with class labels to participate in achieving the best classification results.

(4) The dimension of features is reduced but those features that have no relation with a class label are eliminated.

3. Proposed methodology

In this article, the proposed methodology consists of different stages. The first stage is data gathering. In the first stage, data is downloaded from the UCI database called spambase UCI. The second step is to normalize all the attributes of the dataset which have a higher range of values. In the third step, the feature selection technique is applied which is called Point-Biserial correlation. After that in the fourth step, eight machine learning classifiers are applied to selected attributes as shown in Figure 1.

The proposed methodology, the environment of hardware and software was set as needed to perform experiments. The hp laptop core i5 4th generation having 8 GB RAM is used for experimentation. The PYCHARM software is used which is the Integrated Development Environment for the python language in which we programmed our experiments. All the latest libraries of python are used for experiments like NumPy, Sklearn and Stratified K-Fold.

3.1 Email dataset

Spambase UCI dataset is used in this article which is downloaded from UCI machine learning repository. This dataset includes 57 attributes having continuous and discrete values. It consists of 4601 instances with given labels in the first instance. In the last column of the dataset, a class label is given which consists of 1 and 0 values. 1 means email is spam and 0 means email is not spam. Most attributes of the dataset indicate the occurrence of a particular word and some special characters in the email. The last three attributes indicate the longest, average and total capital letter sequences length in email texts.

3.2 Preprocessing

Most datasets available on the Internet are not preprocessed. The definition of spambase UCI attributes is given in Table 2.

The spambase UCI dataset attributes have many value ranges, this large range of value normalization technique is applied by Eqn (1).

$$x_{\text{normailized}} = \frac{x - x_{\text{minimum}}}{x_{\text{maximimu}} - x_{\text{minimum}}} \tag{1}$$

The instance value of the specific attribute is denoted by x. x_{minimum} is the minimum value and x_{maximum} is the maximum value in a specific attribute which is to be normalized. $x_{\text{normalized}}$ is the normalized value.

3.3 Feature selection

To select the best attributes from a list of attributes, Point-Biserial correlation coefficient [17] is applied. Point-Biserial correlation is applied where one attribute value is continuous and

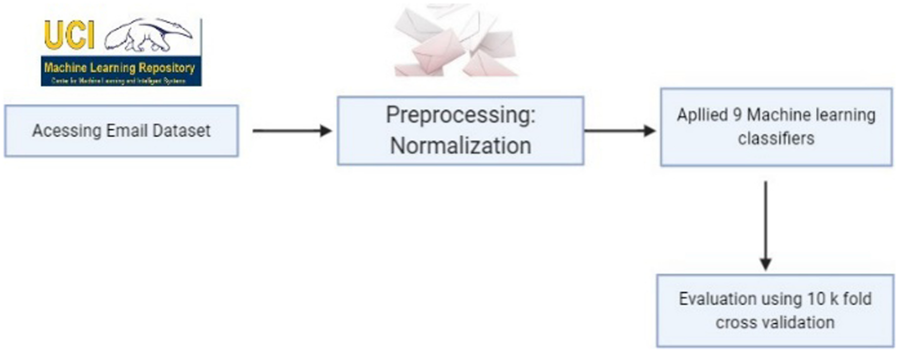
UCI
Machine Learning Repository
Acessing Email Dataset
Preprocessing: Normalization
Apllied 9 Machine learning classifiers
Evaluation using 10 k fold cross validation

Figure 1. Proposed methodology steps for spambase UCI dataset

Attributes	Data type	Description
48	Continuous Real	Percentage of words in the email that match WORD, i.e. 100*(Number of times the WORD appears in the email)/Total number of words in the email. A “word” in this case is any string of alphanumeric characters bounded by non-alphanumeric alphanumeric characters or end of string
06	Continuous Real	Percentage of characters in the email that match CHAR, i.e. 100*(Number of CHAR occurrences)/Total characters in email
01	Continuous Real	The average length of uninterrupted sequences of capital letters
01	Continuous Real	Length of the longest uninterrupted sequence of capital letters
01	Continuous Real	The sum of the length of uninterrupted sequences of capital letters = Total number of capital letters in the email
01	Nominal	Denotes whether the email was considered spam (1) or not (0), i.e. unsolicited commercial email

Table 2.
Description of attributes used in spambase UCI dataset

another value is to be dichotomous. Dichotomous is also known as a binary value. The point-Biserial correlation coefficient is calculated by Eqn (2).

$$r_{pb} = \frac{(M1 - M0)}{Sn} \sqrt{\frac{n1n0}{n^2}} \tag{2}$$

To calculate r_{pb}, dichotomous variables divide into two groups 1 and 0. M1 is the mean value of all the data points which lie in group 1 and M0 is the mean value of all the data points which lie in group 0. $n1$ is several data points in group 1 and n0 is several data points in group 0. At last, n is the total sample size. Point-Biserial correlation is applied to each attribute concerning the class label. Those attributes are not selected whose r_{pb} value is equal to 0. The data used in this research fulfill the requirement of this feature selection technique due to values of features which are continuous and dichotomous class labels. This is the first time we are using it in this domain of email classification according to the best knowledge got from the literature, there is no prior use of it in the domain of spam email classification.

3.4 Classification techniques

The process in which items are combined is based on the similarity between data and the definition of a group. Machine learning classifiers play an important role to classify a large amount of data. In this article, we use different types of machine learning classifiers to predict class labels using the spambase email dataset. The dataset is split into two parts: one is training and the other is testing with the ratio of 70 and 30 size. The training dataset is used to train classifiers model and the testing dataset is used to test the trained model. The classifiers applied in this article are Naive Bayes, Random Forest, K-Nearest Neighbor, Radial Basis Function, Decision Tree, Artificial Neural Network, Logistic Regression and Support Vector Classifier.

Naive Bayes classifier is built for phishing email filtering in Microsoft [2]. It is based on probability and is used to solve classification problems. The training dataset is given to the Naive Bayes model to train the model. Naive Bayes is calculated using Eqn (3).

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \tag{3}$$

where A is the class label and B is the attribute. *P(A/B)* is the posterior probability of class label(A) given attribute(B). *P(B/A)* is the likelihood which is the probability of attribute given class label. *P(A)* is the prior probability of class label and *P(B)* is the prior probability of attribute.

One of the algorithm for classification is RF. Bremen introduced a classifier in 2001 called RF. It consists of multiple decision trees to predict the class label. It is used for classification as well as regression problems. It is very effective against noise and outliers in data and it deals with thousands of inputs without any deletion. RF is measured on classification data using the Gini index. The Gini index formula is defined by Eqn (4).

$$\text{Gini} = \sum_{i=1}^{C} (pi)^2 \tag{4}$$

Gini index uses class and probability to define the Gini of each branch. *C* indicates the number of classes. *pi* indicates the relative frequency of the class. The training part of the dataset is used to train the model of RF and the testing part of the dataset is used to test the trained model of RF.

RBF is a part of the ANN. As compared to multiple hidden layers network, RBF computing speed is fast. It has many uses like classification, time series prediction and system control. In the RBF, every hidden node which is present represents one of the kernel functions. Each kernel function range is defined by its center and width. When attributes are near to center, it means output of kernel function is high and the output of kernel function is reduced to zero as attributes' distance starts to increase from zero. One of the popular kernel functions is the Gaussian function which is applied to training data to train the model of the RBF and testing data to test the trained model of RBF. The RBF consists of inputs, hidden layer and output. Mathematically, input $F_k(x)$ to the *k*th output node is given by

$$f_k(x) = \sum_{q=1}^{Q^k} h_q^k G_q^k(x) \tag{5}$$

Q^k represents the number of hidden nodes linked with target *k*, *q* refers to q^{th} target *k* hidden node and $G^k{}_q(x)$ is the response function of the q^{th} hidden node for target *k*.

The Decision Tree is a graphical representation of possible solutions. It is predictive model learning and it is used for the classification to predict the categorical class label. It works to build a tree that represents different rules for classifying class labels. It considers all attributes to be equally important and independent. The top node in the tree is the known root node and the last nodes are leaf nodes. The Decision Tree can handle both categorical and numerical data. To draw a Decision Tree firstly, we find the entropy of the complete Decision Tree. Secondly, we find the information gain of every attribute. The attribute which has the greatest information gain will be chosen. To calculate two types of entropy and information gain, formulas are defined in Eqs (6) and (7).

$$E(S) = 1 - \sum_{i=1}^{c} (-pi)\log 2(pi) \tag{6}$$

where Eqn (6) represents the frequency table of one attribute. *S* represents the target attribute or class. *Pi* is the frequent probability of element or class in our data.

$$E(T, X) = \sum_{ceX} P(c)E(c) \tag{7}$$

where entropy is defined by the frequency table of two attributes. *T* is the target label and *X* is the attribute. *E(c)* is the entropy of the attribute and *P(c)* is the attribute probability.

$$\text{Gini} = 1 - \sum_{i=1}^{C} (pi)^2 \tag{8}$$

Gini index uses class and probability to define the Gini of each branch. C indicates the number of classes. pi indicates the relative frequency of the class. The training part of the dataset is used to train the model of RF and the testing part of the dataset is used to test the trained model of RF.

ANN is an important classifier in machine learning algorithms. It consists of three layers. The first layer is called the input layer which takes all attributes of the data. The first layer size depends on the number of attributes in the data. The second layer is the hidden layer and its size depends on results taken from multiple experiments. The third layer is the output layer and its size depends on class label values of data. The Multilayer Perceptron is applied to data whose parameters are two hidden layers. Each hidden layer consists of five nodes and the alpha learning rate is 0.01. Boyden–Fletcher–Goldfarb–Shannon is an activation function that is applied in Multilayer Perceptron. Randomly weights are assigned and multiplied with each attribute value. Sum all the product values of attributes and weights. After that activation function is applied on summation and supplied toward the output layer. Weight new formula is written below in Eqn (9).

$$\text{weight}_{\text{new}} = \text{weight}_{\text{old}} + a * (\text{expected} - \text{predicted}) * x \tag{9}$$

where $\text{weight}_{\text{old}}$ is old weight and α is the learning rate. x is the attribute value of data.

LR is used for biological sciences in the early years. It is used in classification problems where the target variable is categorical. In logistic regression a specific threshold is defined. One class is considered above the threshold and below the threshold another class is considered. Its graph shape is just like the S shape. Its value strictly ranges between 1 and 0. The Sigmoid activation function is used in logistic regression. The equation of the linear model is defined by Eqn (10).

$$y = b_0 + b_{1x} \tag{10}$$

where y is the predicted value which depends on x and x is the independent value. b_0 and b_1 are constants. b_0 moves the curves left or right and b_1 is the steepness of the curve. LR is calculated as follows:

$$p = \frac{1}{1 + e^{-(b0+b1x)}} \tag{11}$$

where e is exponential whose value is equal to 2.7182 and the above equation represents the Sigmoid function on which logistic regression depends.

Support vector classifier is a supervised machine learning algorithm. It predicts class labels by maximizing the distance between classes which is called a hyperplane. The vectors which define the hyperplane are called support vectors. It efficiently separated linear and nonlinear attributes. To optimize the result the formula of minimizing $w2 = w^T w$ is calculated as follows.

$$\min_{wbc} \frac{1}{2} w^t w + C \sum_{i=1}^{c} \zeta in \tag{12}$$

where C is the penalty term that controls the strength. Some samples should be at a distance ζi from their correct margin boundary.

KNN is the supervised machine learning algorithm that can be used in classification as well as for regression problems. It stores all training samples and predicts testing samples based on distance function. The classification of testing data is based on most neighbor votes. The distance function is measured by Eqn (13).

$$d = \sqrt{\sum_{i=1}^{k} (xi - yi)^2} \tag{13}$$

where d is equal to the Euclidean distance function. This equation is only valid for continuous variables. If $k = 1$, then the closest class neighbor will be assigned to the testing case. There are several advantages of KNN given as follows:

(1) No need for the additional parameter to add in the KNN model.

(2) It is easy to use and implement.

(3) It is a versatile model that can be used for multiple purposes.

The training set of data is given to the KNN model to train the KNN model and the testing set is given to the KNN trained model to predict.

4. Evaluation and results

In the end, the performance of all the machine learning classifiers applied, such as Naive Bayes, Random Forest, Radial Basis Function, Decision Tree, Artificial Neural Network, Logistic Regression, Support Vector Classifier and K-Nearest Neighbor, is evaluated. Evaluation is done using a confusion matrix table and by calculating Precision, Recall, Accuracy and *F*-Measure of each classifier. Precision, Recall, Accuracy and *F*-Measure are validated using 10-fold cross-validation. 10-fold cross-validation is the splitting of whole data into 10 different parts and 10 iterations are performed on all data. In the first iteration, the first part of data will become test data and all other 9 parts of data will become train data. In the second iteration, the second part of data will become test data and all other parts from 1 to 10 except 2 parts will be train data and so on till 10 iterations. The confusion matrix is described in Figure 2. Figure 2 explains the table of confusion matrix which consists of the following things:

(1) True Positive (TP): Total number of emails predicted spam those in actual are spam emails.

(2) False Positive (FP): Total number of emails predicted spam those in actual are ham emails.

(3) False Negative (FN): Total number of emails predicted ham those in actual are spam emails.

(4) True Negative (TN): Total number of emails predicted ham those in actual are ham emails.

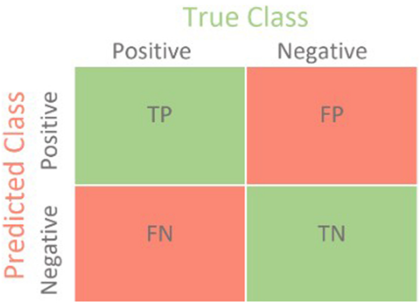

Figure 2.
Diagram of confusion matrix

Four measurements are used for the evaluation of machine learning classifiers: Precision, Recall Accuracy and *F*-Measure. The formulas are given below in the equations, respectively.

$$\text{Precison} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{14}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{15}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \tag{16}$$

The measured average values of TP, FP, TN and FN of confusion matrix for each classifier at 10-fold cross-validation are defined in Table 3. This table shows the minimum value of FP and FN found in ANN that is Multilayer Perceptron with the parameter of two hidden layers and both hidden layers have five nodes. The learning rate in Multilayer Perceptron is α which is equal to 0.01. Table 4 shows the evaluation measures: Precision, Recall, Accuracy and *F*-Measure to calculate the performance of each classifier. The highest accuracy achieved by ANN is 0.9806. Figure 3 shows the graph about the performance of all the machine learning classifiers by using the evaluation measures Precision, Recall, Accuracy and *F*-Measure.

4.1 Comparison of results

The Point-Biserial correlation is used to measure the relationship between a continuous variable and a binary variable. The reason behind the use of this feature selection is that it supports the dataset used in this research to have continuous value in input features and the class label is binary value. The spambase UCI emails dataset has been used by previous papers and their experiment results are compared with implemented experiment

Machine learning classifiers	True positive	False positive	True negative	False negative
Naive Bayes	165	9	269	16
Random Forest	172	13	264	9
Radial Basis Function	166	12	266	15
Decision Tree	199	13	271	9
Artificial Neural Network	178	5	273	3
Logistic Regression	174	3	275	7
Support Vector Classifier	175	3	275	6
K-Nearest Neighbor	154	15	263	26

Table 3.
Results about confusion matrix of each classier at 10-fold cross-validation

Machine learning classifiers	Precision	Recall	Accuracy	*F*-measure
Naive Bayes	0.9482	0.9116	0.9447	0.9295
Random Forest	0.9297	0.9502	0.9500	0.9398
Radial Basis Function	0.9325	0.9171	0.9402	0.9247
Decision Tree	0.9521	0.8361	0.9706	0.8903
Artificial Neural Network	0.9726	0.9834	0.9806	0.9780
Logistic Regression	0.9830	0.9613	0.9771	0.9720
Support Vector Classifier	0.9831	0.9668	0.9800	0.9749
K-Nearest Neighbor	0.9112	0.8555	0.9087	0.8825

Table 4.
Evaluation measures of each classifier on spambase UCI classification

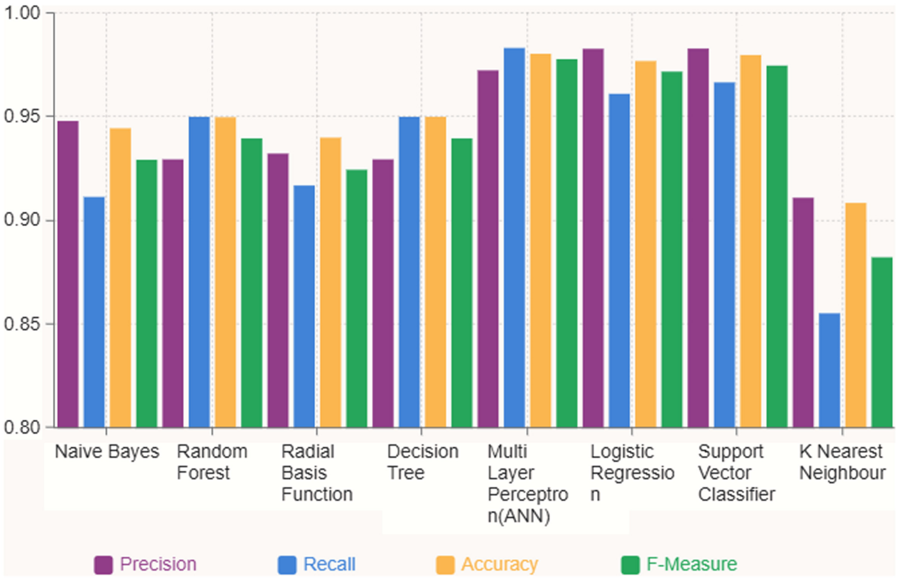

Figure 3. Classifier's performance

(see Table 5). In [4], a total of ten machine learning classifiers are applied to the dataset and the highest accuracy is achieved by using Infinite latent feature selection with the RF that is 95.45%. In [15], the accuracy achieved is 85.31% by using weighted feature selection with ANN on spambase UCI dataset. The accuracy was achieved better by using the Point-Biserial feature selection because it helps us to extract the relevant features for the classification of spam and ham emails. Our proposed method conclude that the highest accuracy is achieved by using the Point-Biserial feature selection and the selected features are used as input for ANN which achieves the accuracy of 0.9806%.

The performance measure of classifiers that give the best result in Table 4 are ANN and Support Vector Machine. In Table 6, we show the comparison of SVM and ANN with Point-Biserial features selection and without Point-Biserial feature selection.

Method	Machine learning classifier	Accuracy achieved (%)
Infinite latent feature selection	Random Forest	95.45
Weighted feature	ANN	85.31
	Logistic Regression	0.9771
Point-Biserial feature selection	Support Vector Classifier	0.9800
	Artificial Neural Network	0.9806

Table 5. Comparison of results on spambase UCI emails dataset

	Classifier	Accuracy (%)
With Point-Biserial	Support Vector Machine	98
With Point-Biserial	Artificial Neural Network	98.06
Without Point-Biserial	Support Vector Machine	89
Without Point-Biserial	Artificial Neural Network	97

Table 6. Comparison without Point-Biserial feature selection and with Point-Biserial feature selection

The result in Table 6 concludes that the accuracy gets higher by using the Point-Biserial correlation feature selection.

5. Conclusion

Different articles use different types of feature selection techniques and different machine learning classifiers to achieve the best results. We use the spambase UCI dataset of 4601 emails. Experiment was performed using different machine learning classifiers which are NB, RF, KNN, RBF, DT, ANN, LR and SVM. We evaluate the performance of all the classifiers and get the highest accuracy of 98.06% using the Multilayer Perceptron classifier. We use the feature selection technique which is Point-Biserial selection on each classifier to select the best features from the rest of all. The classifiers Naive Bayes, Random Forest, K-Nearest Neighbor, Radial Basis Function, Decision Tree, Artificial Neural Network, Logistic Regression and Support Vector Classifier achieved accuracy of 94.47, 95.00, 90.87, 94.02, 95.02, 98.06, 97.71 and 98%, respectively. The best performance is achieved by ANN (Multilayer Perceptron) which obtained an accuracy of 98.06%. We select the best features in our model using Point-Biserial feature selection and applied traditional machine learning algorithms.

We have found out the best features from the listed features but in future work, the new features can be proposed to achieve a better result and the dataset can be increased for using advanced deep learning models.

References

1. Manisha AS, Manisha RJ. Data pre-processing in spam detection. Int J Sci Technol Eng. 2015; 1(11): 5.
2. Bhat SY, Abulaish M, Mirza AA. Spammer classification using ensemble methods over structural social network features. 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). Vol. 2. IEEE, 2014.
3. Trivedi SK, Dey S. Interplay between probabilistic classifiers and boosting algorithms for detecting complex unsolicited emails. J Adv Comp Netw. 2013; 1(2): 132-6.
4. Bassiouni M, Ali M, El-Dahshan EA. Ham and spam e-mails classification using machine learning techniques. J Appl Security Res. 2018; 13(3): 315-31.
5. Mohammad R, Mustafa A. A lifelong spam emails classification model. Appl Comput Inform. 2020; 18(1/2). doi: 10.1016/j.aci.2020.01.002.
6. Ravi Kiran SS, Atmosukarto S. Spam or not spam–that is the question. Citeseerx. 2009.
7. Awad WA., ELseuofi SM. Machine learning methods for E-mail Classification. Int J Computer Appl. 2011; 16(1): 39-45.
8. Alsmadi I, Alhami I. Clustering and classification of email contents. J King Saud University-Computer Inf Sci. 2015; 27(1): 46-57.
9. Sharma S, Amit A. Adaptive approach for spam detection. Inter National J Computer Sci Issues (IJCSI). 2013; 10(4): 23.
10. Chen, XL, *et al.* A method of spam filtering based on weighted support vector machines. 2009 IEEE International Symposium on IT in Medicine & Education. Vol. 1. IEEE, 2009.
11. Scholar M. Supervised learning approach for spam classification analysis using data mining tools. organization. 2010; 2(8): 2760-6.
12. Youn S, McLeod D. A comparative study for email classification. Advances and innovations in systems, computing sciences and software engineering. Dordrecht: Springer; 2007. p. 387-91.
13. Provost J. Naıve-bayes vs. rule-learning in classification of email. Princeton, NJ: Citeseerx; 1999.

14. Bhat SY, Abulaish M, Mirza AA. Spammer classification using ensemble methods over structural social network features. 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). Vol. 2. IEEE, 2014.

15. Alghoul A, *et al*. Email classification using artificial neural network. IJAER; 2018. 2. 8-14.

16. Kaur G, Neelam O. A review article on Naive Bayes classifier with various smoothing techniques. Int J Computer Sci Mobile Comput. 2014; 3(10): 864-8.

17. Tate RF. Correlation between discrete and a continuous variable. Point-biserial correlation. Tate Source : The Ann Math Stat. 1954; 25(3): 603-7.

Corresponding author
Khalid Iqbal can be contacted at: khalidiqbal@cuiatk.edu.pk