

ITC 3/50 Information Technology and Control Vol. 50 / No. 3 / 2021 pp. 411-423 DOI 10.5755/j01.itc.50.3.27349	Development of Proposed Ensemble Model for Spam e-mail Classification	
	Received 2020/07/18	Accepted after revision 2021/06/08
	 http://dx.doi.org/10.5755/j01.itc.50.3.27349	

HOW TO CITE: Shrivastava, A. K., Dewangan, A. K., Ghosh, S. M., Singh, D. (2021). Development of Proposed Ensemble Model for Spam e-mail Classification. *Information Technology and Control*, 50(3), 411-423. <https://doi.org/10.5755/j01.itc.50.3.27349>

Development of Proposed Ensemble Model for Spam e-mail Classification

Akhilesh Kumar Shrivastava

Guru Ghasidas Vishwavidyalaya, Bilaspur (C.G.), India

Amit Kumar Dewangan

Dr. C. V. Raman University, Bilaspur (C.G.), India

S. M. Ghosh

Dr. C. V. Raman University, Bilaspur (C.G.), India

Devendra Singh

Guru Ghasidas Vishwavidyalaya, Bilaspur (C.G.), India

Corresponding author: akhilesh.mca29@gmail.com

Spam e-mail documents classification is a very challenging task for e-mail users, especially non IT users. Billions of people using the internet and face the problem of spam e-mails. The automatic identification and classification of spam e-mails help to reduce the problem of e-mail users in managing a large amount of e-mails. This work aims to do a significant contribution by building a robust model for classification of spam e-mail documents using data mining techniques. In this paper, we use Enron1 data set which consists of spam and ham documents collected from Kaggle repository. We propose an Ensemble Model-1 that is an ensemble of Multilayer Perceptron (MLP), Naïve Bayes and Random Forest (RF) to obtain better accuracy for the classification of spam and ham e-mail documents. Experimental results reveal that the proposed Ensemble Model-1 outperforms other existing classifiers as well as other proposed ensemble models in terms of classification accuracy. The suggested and proposed Ensemble Model-1 produces a high accuracy of 97.25% for classification of spam e-mail documents.

KEYWORDS: Ensemble Model, Classification, Data Mining, Spam e-mail, Machine Learning.

1. Introduction

Many of the previous research work on data mining have focused on structured data. However, in fact, text databases store a valuable section of available information. The text database is a collection of huge amount of documents collected from various sources like news stories, books, digital library, research papers, e-mails, web pages and various social media sites.

These days, a vast majority of data in government, industry, business, and different organizations are put away electronically, as text databases [23]. The entire world is using new technologies for communicating all over the world where e-mail is one of the significant and fast communication media through which we can share information from one e-mail user to another. The main reason why spam e-mails are continuously increasing in mailbox is lack of awareness among the Internet users. Due to this problem, the spam e-mail text (documents) classification is of significance in research work.

Various reputed labs generated report of spam e-mails of every quarter to create awareness in every Internet user. According to Kaspersky Lab report in the first quarter (Q1, 2018) [44], the highest source of spam generating country was Vietnam with 9% spam e-mails, while India was in the 4th position with 7.1% and the average percentage of spam in global e-mail traffic was 51.82%. In the second quarter (Q2, 2018) [44], the highest source of spam generating country was China with 14.36% spam e-mails, while India was in the 11th position with 2.11% and the percentage of spam e-mail traffic in the world was 49.66%. In the third quarter (Q3, 2018) [44], the highest source of spam generating country was China with 13.47% spam e-mails, while India was in the 9th position with 2.84% and the percentage of spam e-mail traffic globally was 52.54%. According to Kaspersky Lab report for the first quarter (Q1, 2019) [44], the highest source of spam was China with 15% spam e-mails, while India was in the 9th position with 2% and the average percentage of spam in the global e-mail traffic was 55.97%.

Spam e-mail is garbage e-mail sent by spammers for their own true intension. These immense quantities of spam e-mails are making a major issue as far as correspondence data transmission use, extra space in mail box and time expended to erase or keep up and maintain.

In a nutshell, this research work contributes the following:

- 1 Pre-process of Enron1 data set.
- 2 Analyse the different individuals and well known ensemble data mining based classification techniques using Enron1 data set.
- 3 Development of the proposed ensemble model based on data mining based classification techniques.
- 4 Comparative analysis with other existing developed models.

The remaining part of this paper is organized as follows: Section 2 explores the review of literature related to spam e-mail classification, Section 3 explores the framework of spam e-mail classification using the proposed method and also explores different methods and materials used in this research work, Section 4 elaborates the experimental results, Section 5 analyses the results and finally Section 6 concludes the research work and also gives the future direction.

2. Related Works

Many researchers have worked in the area of spam e-mail classification using different machine learning techniques and their findings and results are very important to be taken as reference for exploring the new dimension of research work.

Dedeturk and Akay [14] proposed a new spam detection technique through a combination of artificial bee colony algorithm with a logistic regression technique and they also worked on three different datasets to upgrade and handle high-dimensional data with high accuracy. Saidani et al. [32] suggested and used text semantic analysis to improve the performance of model for spam detection. They also suggested automatically extracted semantic features selection technique for spam detection in respective domain. Harisinghaney et al. [17] discussed the detection and classification of text as well as image based e-mail and ham data. They used three classification algorithms namely K-Nearest Neighbors, Naive Bayes and reverse DB-SCAN al-

gorithm for classification of spam e-mails. The performance of these classifiers were evaluated before and after preprocessing of data and produced satisfactory results in terms of accuracy, precision, sensitivity and specificity. Méndez et al. [25] suggested feature selection based semantic ontology to form groups of words for filtering spam e-mails. They used Latent Dirichlet Allocation, information gain, generative statistical model and semantics based feature selection techniques to design spam e-mails filter. Kauret et al. [21] focused on two interlinked problems for representing spam detection and classification. They also explored the various research gaps through this paper for future scope. Dalkilic and Sipahi [13] developed a spam detection model for analyzing the IP address of A and MX records using Sender Policy Framework (SPF) protocol. Barushka and Hajek [5] proposed a novel spam filter approach known as DBB-RDNN. They also compared the performance of proposed spam filtering techniques with different machine learning approaches and achieved better accuracy. Palivalet al. [30] studied the limitations of spam blacklisting system and signature based system and proposed the ID3 algorithm which is based on decision tree technique for spam filtering. The algorithm produced better accuracy compared to the others. Varghese et al. [42] suggested Naïve Bayes classification algorithms using mahout framework to analyse the executing time and accuracy efficiencies. Dada and Joseph [12] used RF machine learning algorithm in WEKA environment. They developed a robust spam e-mail filter with less number of features. Borde et al. [7] used various classification techniques like Naïve Bayes, Perceptron and C4.5 and compared the performance of classifiers for classification of spam and ham documents. They suggested Naïve Bayes classifier which provided a better accuracy over other algorithms. Chouhan [9] used SVM lite tool with four kernel functions for classification of spam e-mails. They also worked on the dataset and calculated different utility function like term frequency (TF), inverse document frequency (IDF) and TF-IDF. They suggested that the SVM classifier is better for classification of spam e-mails and ham e-mails. Dadaand Bassi [11] suggested Logistic Model Tree Induction Algorithm in WEKA environment for classification of spam e-mails filtering and achieved a better accuracy over other conventional techniques. Saleh et al. [33] proposed Negative Selec-

tion Algorithm for identification and classification of spam e-mails. The proposed method gave the highest accuracy of 93.14% with the Enron1 spam e-mail data set. Diale et al. [15] proposed a novel feature extraction and feature dimension reduction techniques to reduce the space complexity and computationally increase the performance of classifiers like SVM, RF and C4.5 decision tree for classification of spam e-mails. Bahgat et al. [4] suggested Word Net ontology, semantic based methods and similarity measures for reducing the extracted textual features, reducing the space and time complexities. The Principal Component Analysis (PCA) and Correlation Feature Selection (CFS) were used to reduce space complexity and semantic filtering approach combined with the feature selection techniques which achieved high computational performance. Ordás et al. [29] developed Concept Drift Analyzer tool for recognizing the ham and spam e-mails with high accuracy using the K-fold cross-validation technique. Naveiro et al. [28] analysed adversarial risk classification using Naïve Bayes algorithm and ACRA framework approach. Basto-Fernandes et al. [6] suggested the rule based multi-objective optimization problem which is an extension version of anti-spam filtering. Yu et al. (2020)[47] proposed a new technique for generating new phishing e-mail data that can be used to train the classifier with high quality data. Venkatraman et al. [43] proposed the integration of Naïve Bayes (NB) with conceptual and semantic similarity technique for classification of spam e-mails. Dada et al. [10] discussed various machine learning techniques for spam e-mails classification in a systematic way. This research work covered a survey and examined the application of machine learning techniques in the context of spam e-mails classification with different spam e-mail datasets. Mohammad [27] proposed a novel model called ELCADP for a lifelong spam e-mails classification. This model was developed for the classification of spam e-mail documents and compared the performance with other techniques, where the proposed model gave better results. Yu et al. [48] proposed a novel spam filtering analyser for generating new spam samples, hence, the spam filtering analyser was able to increase the generalization of classifier. Hota et al. [18] proposed a novel Remove Replacement Feature Selection Technique (RRFST) along with two decision tree techniques for the classification of phishing e-mails.

The above literature review reveals that identification and classification is a very challenging task. It also emphasizes that the strength of the existing classification techniques can be utilized to develop new models. Most of the researchers have emphasized more on classification with feature selection techniques. These literatures help to contribute toward the development of a new ensemble model empowering e-mail users to protect information from unauthorized persons.

3. A Framework of Spam e-mail Documents Classification

This research work proposes an ensemble model for classification of spam and ham e-mail documents. The proposed ensemble model is developed using a combination of different data mining based classification techniques to achieve better classification accuracy. In this architecture, we firstly pre-process the spam and ham e-mail documents and group the different folders of spam and ham e-mails documents into a single folder. Then, the spam e-mail dataset is divided into training and testing data partition using 10-fold cross validation. We input the training and testing dataset into different individuals as well as ensemble classifiers. The proposed new ensemble model is a combination of different individual classifiers. The main motive of the proposed ensemble model is to achieve better classification accuracy compared to each individual classifiers. In this research work, we propose four ensemble models namely Ensemble Model-1, Ensemble Model-2, Ensemble Model-3 and Ensemble Model-4, where Ensemble Model-1 is a combination of MLP, NB and RF, Ensemble Model-2 is a combination of MLP, NB and SVM, Ensemble Model-3 is a combination of SVM, NB and RF and Ensemble Model-4 is a combination of MLP, NB, RF and SVM. Finally, we compare the performance of the proposed ensemble models with different individuals as well as existing ensemble classifiers in measures of accuracy, sensitivity, specificity, precision, F-score and ROC curve, where Ensemble Model-1 gives a better performance compared to the others. Figure 1 shows the flow of the proposed work for classification of spam e-mail documents. The pseudo code of the proposed model is given below:

Pseudo Code of Proposed Model

Input

Hset={Set of Ham e-mails}

Sset={Set of Spam e-mails}

HSset={Hset, Sset}={Set of Ham and Spam e-mails}

Output

PM = Performance Measures={Ac, Sen, Spc, Pr, Fs, roc, auc}

M_{best} = Best model

Where Ac= Accuracy, Sen= Sensitivity, Spec= Specificity, Pr= Precision, Fs=F-score,

roc = receiver operating characteristic (ROC),

auc= area under curve (AUC)

Ensemble Model (HSset, M_1 , M_2 , M_{best})

1. Start
2. Apply HSset dataset to different individuals and ensemble classifiers.

M_1 =HSset→{Gaussian Naïve Bayes, DT, KNN, MLP, RF, Bagging, AdaBoosting, Gradient Boosting}

M_2 =HSset→{Ensemble Model-1, Ensemble Model-2, Ensemble Model-3, Ensemble Model-4}

M_{best} =Compare{ (M_1 ←Ac), (M_2 ←Ac)}

M_{best} ={ Ac, Sen, Spec, Pr, Fs, roc, auc}

3. End.

3.1. Enron1 Data Set

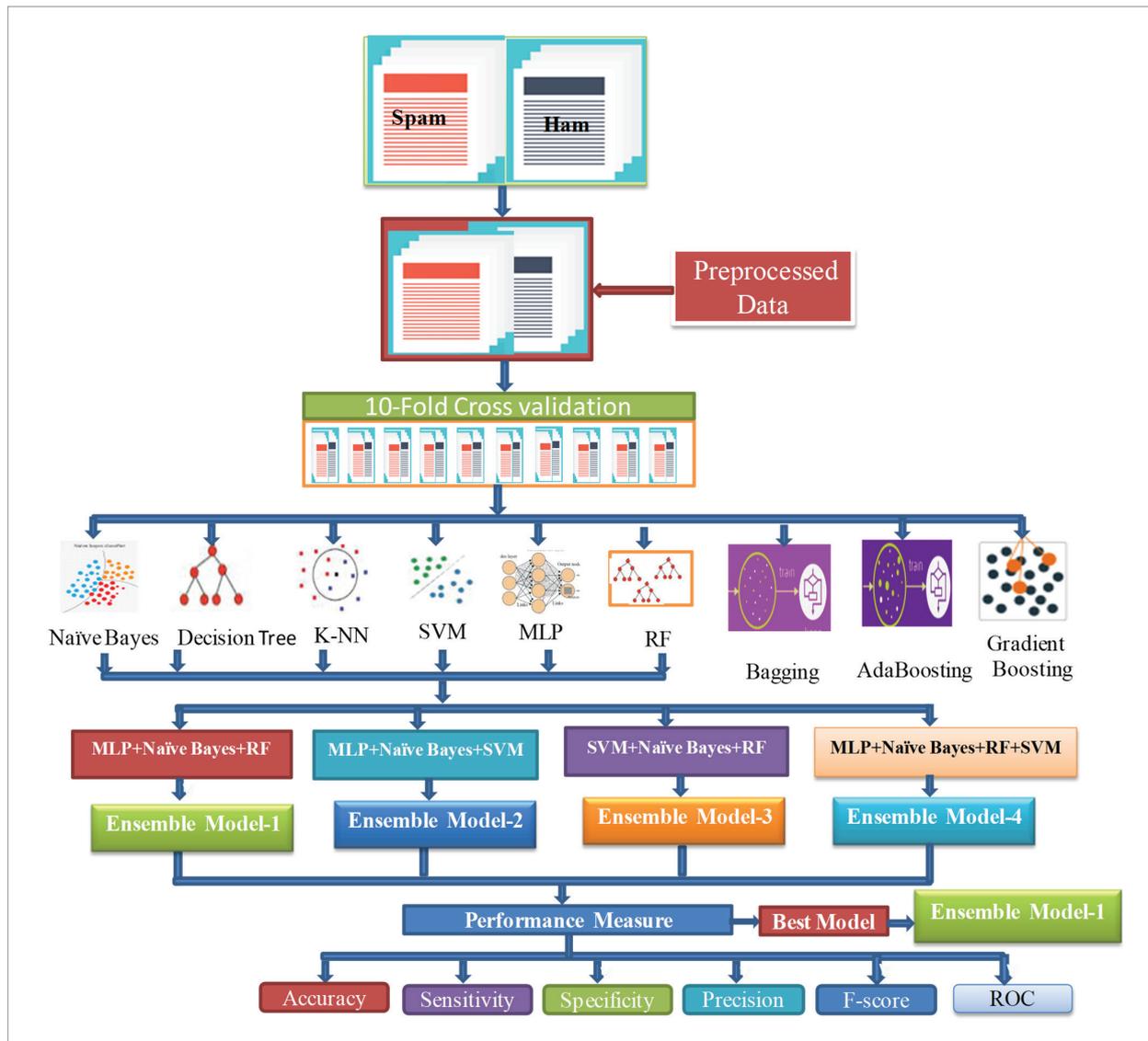
The Enron1 dataset is a collection of spam and ham documents collected from Kaggle repository [45]. This dataset consists of 5975 spam and ham e-mail documents where 1500 e-mails documents belong to spam e-mails while 3672 e-mails belong to ham e-mail documents.

3.2. Cross Validation

K-fold [23] cross validation is a commonly used to evaluate the performance of machine learning techniques. K-fold cross validation is a process of random partition of data into k consecutive folds. In this research work, we performed a K-fold cross validation with k=10 where the dataset was divided 10 times into 10 different training sets (90% of total dataset) and testing sets (10% of total datasets).

Figure 1

Flow of proposed work for classification of Spam e-mail



3.3. Machine Learning Techniques

Machine learning (ML) [46] is a subset of artificial intelligence which is concerned with learning from data, analysing data and get some relevant knowledge from large amount of dataset. The main aim of ML technique is to design and develop robust model which can be used to arrive at the data with better performance. ML can be categorized into supervised, semi-supervised, unsupervised and reinforcement

learning. This research work has used the supervised learning algorithm for classification of spam and ham e-mail documents. Various supervised machine learning techniques used in this research work are discussed below:

3.3.1. Decision Tree (DT)

Decision tree [35-8] is one of the most popular and well-known data mining based classification techniques for classification and prediction task. Each

node in decision tree indicates either a decision node or a leaf node where leaf node represents the value of target attributes of instances. A decision tree is to split dataset into different subsets recursively so that each subset contains more or less homogeneous states of our target variable.

3.3.2. Naïve Bayes (NB)

Naïve Bayes algorithm [16-34] is simple and based on probability theorem. It [23] is a statistical classifier also known as the Naive Bayesian classifier that can be used for classification of data. The performance of Naïve Bayes classifier is to be compared with neural network and decision tree classifier. Bayesian classifiers have also revealed better performance with large amount of databases.

3.3.3. K-Nearest Neighbor (K-NN)

K-NN [23-20] is a data mining technique that is widely used in the field of classification, prediction and pattern recognition. It is also a type of supervised machine learning technique where model is trained with training samples and the trained model is tested with testing samples. Each training sample is described by number of attributes and each sample represents a point in the n-dimensional space.

3.3.4. Support Vector Machine (SVM)

SVM [49] is a supervised learning technique that is useful for solving the traditional classification problem. where each input tuple is associated with one class label. SVM is used for both linear and nonlinear data classification. SVM is based on the concept of hyper plane and divides the n dimensional space of data into two regions. This hyper plane always maximizes the margin between the two regions. The margin is defined by the longest distance between the examples of the two regions and is computed based on the distance between the closest instances of both regions to the margin, which are called supporting vectors.

3.3.5. Multilayer Perceptron (MLP)

Multilayer Perceptron [36] is an advancement from the straightforward perceptron in which extra shrouded layers are included. It contains more than one hidden layer, so it is called multilayer perceptron. MLP structure is formed from the input layer to the first hidden layer, from the first hidden layer to the second and so on, to the output layer to the last hidden layer. MLP handles the non-linear data. It is a super-

vised machine learning that can be used for classification and prediction.

3.3.6. Ensemble Technique

Ensemble technique [24] is a strategy combining two or more models for improving the accuracy compared to other individual models. The main purpose of the ensemble technique is to expand the accuracy and avoid the drawback of individual models. In this research paper we have used Random Forest (RF), Bagging, and Boosting (AdaBoosting and Gradient Boosting) ensemble methods. We have also used voting scheme for combining data mining based classification techniques.

– Random Forest(RF)

RF [31] is an ensemble classifier that is a combination of many decision trees. The main motive of this ensemble classifier is to achieve better accuracy compared to individuals. RF is basically used with very large training datasets and a very large number of input features. A RF classifier is basically a combination of tens or hundreds of decision trees.

– Bagging and Boosting

Bagging and boosting [24] are two well-known ensemble methods that can be used to combine models. The main aim of using this methods are to improve the performance of the model. Both bagging and boosting can be used for classification as well as prediction. In this research work we have used Bagging, AdaBoosting and Gradient Boosting for classification of spam e-mail documents.

– Voting Scheme

Voting scheme [24] is a meta classifier and the most important ensemble technique to combine any classifier through majority of voting. The final class label is predicted by a majority of the classifiers. The final class label F_j is defined as

$$F_j = \text{mode} \{C_1, C_2, C_3, \dots, C_n\},$$

where $\{C_1, C_2, C_3, \dots, C_n\}$ indicates the individual classifiers that participate in the voting. This research work has used voting scheme to develop a proposed classifier.

3.4 Performance Measures

The performance measures [19-38] play a very important role in checking the robustness of a model.

We have calculated accuracy, sensitivity, specificity, precision, F-score and ROC curve using different parameters of the confusion matrix. The confusion matrix includes parameters like true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Various performance measures like accuracy, sensitivity, specificity, precision, F-score [37] and ROC curve are calculated using the elements of confusion matrix.

Accuracy [38] is one of the important measures to check the performance of any model. It is the ratio between the correctly classified positive and negative samples to the total number of samples as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Sensitivity [38] is also called True positive rate (TPR), hit rate, or recall. It is represented as the ratio of positive correctly classified samples to the total number of positive samples as follows:

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

Specificity [38] is also called True negative rate (TNR), or inverse recall and is expressed as the ratio of the correctly classified negative samples to the total number of negative samples as follows:

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

Precision [19] can be expressed as the rate of instances classified correctly among the results of classifier.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

F-score [19] is the harmonic mean of precision and recall.

$$\text{F-score} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$$

Receiver operating characteristics (ROC) [22] is another important measure to check the performance of a model. ROC curve represents the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (1-specificity) to check the performance of predictive model where TPR represents the y-axis and FPR represents the x-axis. The main concept of ROC curve is to maintain a balance between the true positives, and false positives.

Area under the ROC curve (AUC) [22] is another performance measure to calculate the area under the ROC curve. The AUC score is always bounded between zero and one.

4. Experimental Results

This experiment work is carried out using Python (Jupyter notebook) with Anaconda environment in Window7 operating system. Nowadays, Python is an emerging software tool for web development, scientific computing, image processing, data analysis, machine learning and deep learning. In this research work, we propose an ensemble model and check the robustness and efficiency of the model. Efficiency and robustness of the proposed ensemble model is verified using different performance measures like accuracy, sensitivity, specificity, and precision, F-score, ROC curve and AUC score with Enron1 dataset. Enron1 data set is a collection of spam and ham documents. We propose four ensemble models namely, Ensemble Model-1, Ensemble Model-2, Ensemble Model-3 and Ensemble Model-4 for the classification of spam and ham documents. This research work uses different individuals, well-known ensemble classifiers and proposed ensemble models for classifying spam and ham e-mail documents as shown in Table 1. Table 1 shows the accuracy of individual classifiers, existing ensemble classifiers and proposed ensemble classifiers, where Naïve Bayes (NB) gives the highest

Table 1

Accuracy of individuals and ensemble classifiers

Category of Classifier	Classifier	Accuracy
Individual Classifier	Naïve Bayes	94.82%
	Decision Tree	91.51%
	K-NN	86.66%
	SVM	94.57%
	MLP	96.06%
Existing Ensemble Classifier	RF	95.92%
	Bagging (BaseC=RF)	95.11%
	AdaBoosting	94.88%
	GradientBoosting	92.81%
Proposed Ensemble Model	Ensemble Model-1	97.25%
	Ensemble Model-2	96.83%
	Ensemble Model-3	96.64%
	Ensemble Model-4	97.10%

Table2

Confusion matrix of proposed ensemble models

Actual Vs Predicted	Ensemble Model-1		Ensemble Model-2		Ensemble Model-3		Ensemble Model-4	
	Ham	Spam	Ham	Spam	Ham	Spam	Ham	Spam
Ham	3567	105	3545	127	3540	132	3594	78
Spam	37	1463	37	1463	42	1458	72	1428

Table3

Performance measures of proposed ensemble models

Performance Measures	Ensemble Model-1	Ensemble Model-2	Ensemble Model-3	Ensemble Model-4
Accuracy	97.25%	96.83%	96.64%	97.10%
Sensitivity	97.14%	96.54%	96.41%	97.87%
Specificity	97.53%	97.53%	97.20%	95.20%
Precision	98.97%	98.97%	98.83%	98.03%
F-score	98.05%	97.74%	97.60%	97.95%

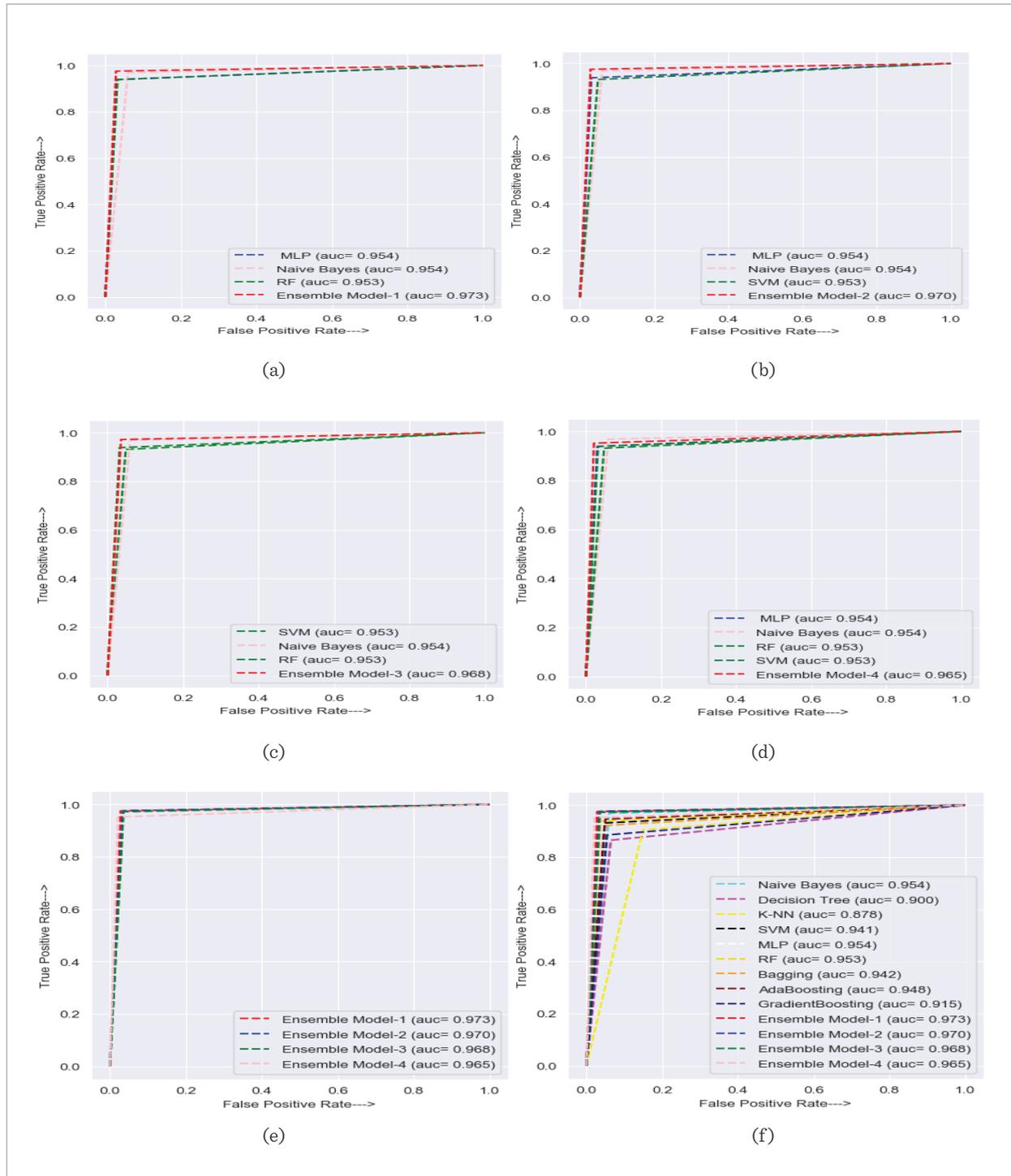
94.82% of accuracy in case of individual classifiers, bagging gives the highest 95.11% of accuracy in the case of existing ensemble classifiers and Ensemble Model-1 gives the highest 97.25% of accuracy in the case of proposed ensemble classifiers. Finally, Table 1 shows that the proposed Ensemble Model-1 is a robust and efficient model for classification of spam e-mail documents. Now, we verify the robustness of the proposed ensemble models using other performance measures like sensitivity, specificity, precision, F-score and ROC curve and Area Under ROC curve (AUC). The performance can be calculated using different parameters of confusion matrix like TP, TN, FP and FN. Table 2 shows the confusion matrix of the proposed Ensemble Model-1, Ensemble Model-2, Ensemble Model-3 and Ensemble Model-4. Table 3 shows the various performance measures of the proposed ensemble models where proposed Ensemble Model-1 gives the highest accuracy of 97.25% compared to other models.

ROC curve [22] represents the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (1-specificity) to check the performance of the predictive model. The curve of the proposed Ensemble Model-1 is closer to the top left corner which indicates the best performance compared to the others.

AUC is an important and effective measure for checking the robustness of models. The maximum value of 1 for AUC means that the model is perfect for classifying the spam and ham samples with zero error. Figure 2 shows the comparative analysis of ROC curve and AUC for individuals and proposed models. Figure 2(a) shows the ROC curve of the proposed Ensemble Model-1 and individual classifiers like MLP, Naïve Bayes and RF, Figure 2(b) shows the ROC curve of the proposed Ensemble Model-2 and individual classifiers like MLP, Naïve Bayes and SVM, Figure 2(c) shows the ROC curve of Ensemble Model-3 and individual classifiers like SVM, Naïve Bayes and RF, Figure 2(d) shows the ROC curve of Ensemble Model-4 and individual classifiers like MLP, Naïve Bayes, RF and SVM. In each comparison, the proposed ensemble models give better accuracy of their individual classifiers. Figure 2(e) shows the comparison of Ensemble Model-1, Ensemble Model-2, Ensemble Model-3 and Ensemble Model-4 in terms of ROC curve, in which the proposed Ensemble Model-1 gives the highest AUC score compared to their individual classifiers. Finally, Figure 2(f) shows the comparative analysis of all the proposed ensemble models and individual classifiers in terms of ROC curve and AUC score where the proposed Ensemble Model-1 gives the highest AUC score

Figure 2

Comparative ROC curve for (a) Proposed Ensemble Model-1 and Individual classifiers, (b) Proposed Ensemble Model-2 and Individual classifiers, (c) Proposed Ensemble Model-3 and Individual classifiers, (d) Ensemble Model-4 and Individual classifiers, (e) Proposed Ensemble Models, and (f) Proposed and Existing Ensemble Models with Individual classifiers



of 0.973 among the other ensemble models as well as individual classifiers. Table 4 shows the comparative analysis of various individual classifiers and the proposed ensemble models with AUC score, in which our proposed Ensemble Model-1 gives the highest AUC score compared to the others. Finally, we conclude that the proposed Ensemble Model-1 is recommended for the classification of spam and ham se-mails.

5. Results Analysis

The proposed Ensemble Model-1 has given better classification accuracy compared to other models previously developed by different researchers on Enron1 dataset, as shown in Table 5. The table below shows that our proposed Ensemble Model-1 is an effective and robust model for the classification of spam and ham e-mails.

Table 4

Comparative analysis of various classifiers with AUC score

Classifier/Proposed Model	AUC Score
SVM	0.953
Naïve Bayes(NB)	0.954
RF	0.953
MLP	0.954
Ensemble Model-1	0.973
Ensemble Model-2	0.970
Ensemble Model-3	0.968
Ensemble Model-4	0.965

Table 5

Comparative analysis of proposed model with previous developed models by different researchers on Enron1 dataset

Author(s)	Technique Used	Accuracy
Abi-Haidar and Rocha [1]	Adaptive Immune System (AIS)	90.00%
Almeida et al. [2]	Multivariate Bernoulli Naive Bayes	94.79%
Almeida and Yamakami [3]	Basic Naïve Bayes	92.86%
Uysal and Gunal [41]	Distinguishing Feature Selection	94.35%
Mishra and Thakur [26]	Random Forest	96.39%
Trivedi and Dey [40]	ReliefF+NB	96.30%
Trivedi and Dey [39]	SVM+ Boosted Naïve Bayes	95.60%
Varghese et al. [42]	Naïve Bayes	93.04%
Borde et al. [7]	Naïve Bayes	91.60%
Bahgat et al. [4]	SVM with Feature selection (CFS) + Semantic relations and similarity measures	94.00%
Saleh et al. [33]	Negative Selection Algorithm (NSA)	93.14%
Naveiro et al. [28]	MC 0.5 ACRA Enron-Spam	82.40%
Mohammad [27]	ELCADP	95.80%
Proposed Ensemble Model-1	MLP, Naïve Bayes and RF	97.25%

6. Conclusions and Future Work

This study has shown that the proposed Ensemble Model-1 outperforms other existing spam e-mail filtering methods in terms of classification accuracy. More importantly, it has classified both spam and ham documents in satisfactory levels. The comparative analysis of proposed Ensemble Model-1 outperformed previous approaches and existing classifiers on Enron1 dataset. The novelty of the proposed model

is to obtain accurate results for classification of spam and ham e-mail documents. However, a further experiment is needed on the other datasets to show that the proposed model is not limited to classification of spam and ham e-mail documents. In future, the proposed model can be effectively applied on high dimension imbalanced text classification problems like news and social network based data as well as sentimental analysis.

References

1. Abi-Haidar, A., Rocha, L. M. Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics: A Study of Concept Drift. Proceedings of the 11th International Conference on the Simulation and Synthesis of Living Systems, 2008, v1-8.
2. Almeida, T. A., Almeida, J., Yamakami, A. Spam Filtering: How the Dimensionality Reduction Affects the Accuracy of Naive Bayes Classifiers. Journal of Internet Services and Applications, 2011, 1(3), 183-200. <https://doi.org/10.1007/s13174-010-0014-7>
3. Almeida, T. A., Yamakami, A. Compression-based Spam Filter. Special Issue on Security and Communication Networks, 2012. <https://doi.org/10.1002/sec.639> <https://doi.org/10.1002/sec.639>
4. Bahgat, E. M., Rady, S., Gad, W., Moawad, I. F. Efficient E-mail Classification Approach Based on Semantic Methods. Ain Shams Engineering Journal, 2018, 9, 3259-3269. <https://doi.org/10.1016/j.asej.2018.06.001>
5. Barushka, A., Hajek, P. Spam Filtering Using Integrated Distribution-Based Balancing Approach and Regularized Deep Neural Networks. Applied Intelligence, 2018, 48, 3538-3556. <https://doi.org/10.1007/s10489-018-1161-y>
6. Basto-Fernandes, V., Yevseyevab, I., Méndezc, J. R., Zhaod, J., Fdez-Riverola, F., Emmerich, M. T. M. A Spam Filtering Multi-objective Optimization Study Covering Parsimony Maximization and Three-Way Classification. Applied Soft Computing, 2016, 48, 111-123. <https://doi.org/10.1016/j.asoc.2016.06.043>
7. Borde, S., Agrawal, U. M., Bilay, V. S., Dogra, N. M. Supervised Machine Learning Techniques for Spam E-mail Detection. International Journal for Science and Advance Research in Technology, 2017, 3(3), 760-764.
8. Bozkir, A. S., Sezer, E. A. A New Web Based Data Mining Exploration and Reporting Tool for Decision Makers. Artificial Intelligence Research, 2013, 2, 70-89. <https://doi.org/10.5430/air.v2n3p70>
9. Chouhan, S. Behavior Analysis of SVM Based Spam Filtering Using Various Kernel Functions and Data Representations. International Journal of Engineering Research and Technology, 2013, 2(9), 3029-3036.
10. Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., Ajibuwa, O. E. Machine Learning for Email Spam Filtering: Review. Approaches and Open Research Problems, Heliyon, 2019, 5, 1-23. <https://doi.org/10.1016/j.heliyon.2019.e01802>
11. Dada, E. G., Bassi, J. S. Logistic Model Tree Induction Machine Learning Technique for Email Spam Filtering. The Pacific Journal of Science and Technology, 2018, 19(2), 96-102.
12. Dada, E. G., Joseph, S. B. Random Forests Machine Learning Technique for Email Spam Filtering. University of Maiduguri Faculty of Engineering Seminar Series, 2018, 9(1), 29-36.
13. Dalkilic, G., Sipahi, D. Spam Filtering with Sender Authentication Network. Computer Communications, 2017, 98, 72-79. <https://doi.org/10.1016/j.comcom.2016.12.008>
14. Dedetürk, B. K., Akay, B. Spam Filtering Using a Logistic Regression Model Trained by an Artificial Bee Colony Algorithm. Applied Soft Computing Journal, 2020, 91, 1-18. <https://doi.org/10.1016/j.asoc.2020.106229>
15. Diale, M., Celik, T., Walt, C. V. D. Unsupervised Feature Learning for Spam E-mail Filtering. Computers and Electrical Engineering, 2019, 74, 89-104. <https://doi.org/10.1016/j.compeleceng.2019.01.004>

16. Gupta, A., Mohan, K. M., Shidnal, S. Spam Filter Using Naïve Bayesian Technique. *International Journal of Computational Engineering Research*, 2018, 8, 26-32.
17. Harisinghane, A., Dixit, A., Gupta, S., Arora, A. Text and Image Based Spam Email Classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm. 2014 International Conference on Reliability, Optimization and Information Technology, 2014. <https://doi.org/10.1109/ICROIT.2014.6798302>
18. Hota, H. S., Shrivastava, A. K., Hota, R. An Ensemble Model for Detecting Phishing Attack with Proposed Remove-Replace Feature Selection Technique. *Procedia Computer Science*, 2018, 132, 900-907. <https://doi.org/10.1016/j.procs.2018.05.103>
19. Hota, H.S., Sharma, D. K., Shrivastava, A. K. Development of an Efficient Classifier Using Proposed Sensitivity-Based Feature Selection Technique for Intrusion Detection System. *International Journal of Information and Computer Security*, 2018, 10(1), 80-101. <https://doi.org/10.1504/IJICS.2018.089594>
20. Jadhav, S. D., Channe, H. P. Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research*, 2015, 5, 1842-1845. <https://doi.org/10.21275/v5i1.NOV153131>
21. Kaur, R., Singh S., Kumar, H. Rise of Spam and Compromised Accounts in Online Social Networks: A State-of-the-Art Review of Different Combating Approaches. *Journal of Network and Computer Applications*, 2018, 112, 53-88. <https://doi.org/10.1016/j.jnca.2018.03.015>
22. Kumar, R., Indrayan, A. Receiver Operating Characteristic (ROC) Curve for Medical Researchers. *Indian Pediatrics*, 2011, 48, 277-287. <https://doi.org/10.1007/s13312-011-0055-4>
23. Laorden C., Santos, I., Sanz, B., Alvarez, G., Bringas, P. G. Word Sense Disambiguation for Spam Filtering. *Electronic Commerce Research and Applications*, 2012, 11(3), 1-9. <https://doi.org/10.1016/j.eelerap.2011.11.004>
24. Latha, C.B.C, Jeeva, S. C. Improving the Accuracy of Prediction of Heart Disease Risk based on Ensemble Classification Techniques. *Informatics in Medicine Unlocked*, 2019, 16, 1-9. <https://doi.org/10.1016/j.imu.2019.100203>
25. Méndez, J. R., Cotos-Yañez, T. R., Ordás, D. R. A New Semantic-Based Feature Selection Method for Spam Filtering. *Applied Soft Computing Journal*, 2019, 76, 89-104. <https://doi.org/10.1016/j.asoc.2018.12.008>
26. Mishra, R., Thakur, R. S. Analysis of Random Forest and Naïve Bayes for Spam Mail Using Feature Selection Categorization. *International Journal of Computer Applications*, 2013, 80(3), 42-47. <https://doi.org/10.5120/13844-1670>
27. Mohammad, R. M. A. A Lifelong Spam E-mails Classification Model. *Applied Computing and Informatics*, 2020, 1-10. <https://doi.org/10.1016/j.aci.2020.01.002>
28. Naveiro, R., Redondo, A., Insua, D. R., Ruggeri, F. Adversarial Classification: An Adversarial Risk Analysis Approach. *International Journal of Approximate Reasoning*, 2019, 113, 133-148. <https://doi.org/10.1016/j.ijar.2019.07.003>
29. Ordás, D. R., Riverola, F. F., Méndez, J. R. Concept Drift in E-mail Datasets: An Empirical Study with Practical Implications. *Information Sciences*, 2018, 428, 120-135. <https://doi.org/10.1016/j.ins.2017.10.049>
30. Palival, D., Printer, K., Devre, R., Lemos, N. Email Spam Filtering Using Decision Tree Algorithm. *International Journal of Scientific and Engineering Research*, 2018, 9(3), 40-42.
31. Parimala, R., Nallaswamy, R. A Study of Spam E-mail Classification Using Feature Selection Package. *Global Journal of Computer Science and Technology*, 2011, 11. ISSN: 0975-4172.
32. Saidani, N., Adi, K., Allili, M. S. A Semantic-Based Classification Approach for an Enhanced Spam Detection. *Computers and Security*, 2020, 94. <https://doi.org/10.1016/j.cose.2020.101716>
33. Saleh, A. J., Karim, A., Shanmugam, B., Azam, S., Kannoopatti, K., Jonkman, M., Boer, F. D. An Intelligent Spam Detection Model Based on Artificial Immune System. *International Journal of Information*, 2019, 10, 1-17. <https://doi.org/10.3390/info10060209>
34. Salmi, N., Rustam, Z. Naïve Bayes Classifier Models for Predicting the Colon Cancer. 9th Annual Basic Science International Conference 2019, 1-8. <https://doi.org/10.1088/1757-899X/546/5/052068>
35. Shi L., Wang, Q., MA, X., Weng, M., Qiao, H. Spam Email Classification Using Decision Tree. *Journal of Computational Information Systems*, 2012, 8(3), 949-956.
36. Shrivastava, A. K., Ghosh, S. M., Dewangan, A. K. Text Classification of Cornell Movie Data Using Data Mining with Feature Selection. *International Journal of Engineering and Advanced Technology*, 2019, 9, 2950-2955. <https://doi.org/10.35940/ijeat.B2329.129219>
37. Sokolova, M., Japkowicz, N., Szpakowicz, S., Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *Australasian*

- Joint Conference on Artificial Intelligence, Springer, 2006, 1015-1021. https://doi.org/10.1007/11941439_114
38. Tharwat, A. Classification Assessment Methods. *Applied Computing and Informatics*, 2019, 1-38. <https://doi.org/10.1016/j.aci.2018.08.003>
39. Trivedi, S. K., Dey, S. A Comparative Study of Various Supervised Feature Selection Methods for Spam Classification. In: *Proceedings of the 2nd International Conference on Information and Communication Technology for Competitive Strategies*. ACM, 2016. <https://doi.org/10.1145/2905055.2905122>
40. Trivedi, S. K., Dey, S. A Combining Classifiers Approach for Detecting E-mail Spams. *30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 2016, 355-360. <https://doi.org/10.1109/WAINA.2016.127>
41. Uysal, A. K., Gunal, S. A Novel Probabilistic Feature Selection Method for Text Classification. *Knowledge-Based System*, 2012, 36, 226-235. <https://doi.org/10.1016/j.knosys.2012.06.005>
42. Varghese, L., Supriya, M. H., Jacob, K. P. Spam: A Big Data Challenge. *International Journal of Advanced Research in Computer Science*, 2017, 8(1), 195-198.
43. Venkatraman, S., Surendiran, B., Kumar, P. A. R. Spam E-mail Classification for the Internet of Things Environment Using Semantic Similarity Approach. *The Journal of Supercomputing*, 76, 2019, 756-776. <https://doi.org/10.1007/s11227-019-02913-7>
44. Web source. <https://securelist.com/spam-and-phishing-in-q1-2019/90795/> Accessed on May 05, 2020.
45. Web source. <https://www.kaggle.com/wanderfj/enron-spam> Accessed on Dec. 05, 2019.
46. Websource: <https://www.journals.elsevier.com/informatics-in-medicine-unlocked/call-for-papers/machine-learning-for-intelligent-decision-making>. Accessed on June 10, 2020.
47. Yu, G., Fan, W., Huang, W. An Explainable Method of Phishing Emails Generation and Its Application in Machine Learning. *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2020, 1279-1283. <https://doi.org/10.1109/ITNEC48623.2020.9085171>
48. Yu, G., Fan, W., Huang, W., An, J. An Explainable Method of Phishing E-mails Generation and Its Application in Machine Learning. *IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference*, 2020, 1279-1283. <https://doi.org/10.1109/ITNEC48623.2020.9085171>
49. Zhanga, C., Shaob, X., Lia, D. Knowledge-based Support Vector Classification Based on C-SVC. *Procedia Computer Science*, 2013, 17, 1083-1090. <https://doi.org/10.1016/j.procs.2013.05.137>

