IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 26, No. 5, October, 2021, pp. 437-444
Journal homepage: http://iieta.org/journals/isi

Robust Text Classifier for Classification of Spam E-Mail Documents with Feature Selection Technique

Akhilesh Kumar Shrivas[1*], Amit Kumar Dewangan[2], Samrendra Mohan Ghosh[2]

[1] Department of CSIT, Guru Ghasidas Vishwavidyalaya, Bilaspur, Chhattisgarh 495001, India
[2] Computer Science Engineering, Dr. C. V. Raman University, Bilaspur, Chhattisgarh 495001, India

Corresponding Author Email: akhilesh.mca29@gmail.com

https://doi.org/10.18280/isi.260502

Received: 13 September 2021
Accepted: 13 October 2021

Keywords:
spam e-mail, classification, preprocessing, random forest (RF), feature selection technique (FST)

ABSTRACT

E-mails are an effective medium for sending information in various modes like text, audio, video, etc. from one person to another. Spam e-mail is a junk e-mail that unnecessary wastage memory space, wasting time to delete and maintain e-mails in the mailbox. The contribution of this research work is to develop a robust and computational efficient classifier that classifies the spam e-mail and ham e-mail documents. This paper analyzes and validates the spam e-mails documents using different data mining-based classification techniques. The most importance of this research work is to select the best classifier with reduce feature subset of datasets that achieve better accuracy compared to other existing classifiers. We have collected six types of Enron datasets and prepared the last seven Enron datasets that combine all these six Enron datasets. Then, filtering the datasets with the help of the WEKA data mining tool. In the first step, we perform preprocessing the datasets and remove all the irrelevant words from the datasets. We have used different classifiers like Nave Bayes, J48, Random Forest, Random Tree, and Adaboosting to analyze and classify ham and spam e-mails documents. We also compare the performance of the classifier in terms of accuracy where Random Forest gives better accuracy with all seven Enron datasets. Finally, we have used the SymmetricalUncert feature selection technique to make the optimized dataset with a reduced feature subset. The suggested Random Forest classifier gives 98.73% of accuracy with reduced features of Enron datasets.

1. INTRODUCTION

E-mail is an effective and efficient communication medium to personal or official for any organization. One of the types of e-mail is the spam e-mail. Spam e-mail is a junk e-mail that is not necessary to harmful for users but it contains the unwanted details and send to the e-mail users by spammers. The first spam e-mail was generated on 3 may 1978 to several thousands of users on ARPANET sent by Gary Thuerk [1].

Sometimes many organizations or any unauthentic person use mails to sell the products, provide attractive offers related to products, send URLs, or any kind of offer greed to users such mails are called spam e-mails. There is no charge payable to sending e-mails so the person/organization sends the bulk number of e-mails to a different recipient. Spam e-mails are helpful to sell products and steal susceptible information. They use this information to involve users in any criminal activity or information to gain financial transactions. We collect this vital and sensitive information of people through spam e-mails, so spam e-mails are very deadly in today's computer age. Spam e-mail is a severe challenge because computer usage is becoming very fast in people today. But even today, people are not as knowledgeable as they should be, so we must face severe spam e-mail challenges very carefully. According to Kaspersky lab report, average spam e-mail traffic was 46.56%, in Q2, 2021 which is grow up 0.89% against the Q1, 2021 [2]. There are many ways like machine learning and classification to face difficulties like spam e-mails. Many researchers are currently using marching learning and data mining-based classification techniques like decision tree, naive bayes, support vector machine etc. as classifiers for classification of spam e-mail documents. The researchers are also using various evolutionary techniques like genetic algorithm, particle swarm optimization, principle component analysis to reduce the feature space of dataset.

Many researchers proposed various models and techniques to prevent spam e-mails. We analyzed those techniques and proposed a model that uses classification and feature selection techniques (FSTs). Classification is very effective techniques by which we can classify spam e-mails. The classification techniques can quickly point out the spam e-mails documents, and also increase accuracy of model by with FSTs. This work even more effectively and efficiently for reorganization and classification of spam e-mail and ham e-mail documents.

The remaining part of this research work as given where section 2 explores the review of literature related to spam e-mail classification, section 3 examines the framework of spam e-mail documents classification using the proposed algorithm and also explore the different methods and materials have used in this paper, section 4 explores the experimental results, section 5 analyzes the result, and finally section 6 concludes the research work and also gives future direction.

2. LITERATURE REVIEW

The literature review is an important section of a research paper. This section explored the research work done by different authors related to classification of spam e-mails documents. The authors [3] used spam e-mails and websites to study the detection of spam and also recognized the effectiveness of the Negative Selection Algorithm (NSA). NSA gave a high accuracy rate and low error rate. The authors [4, 5] used a number of papers related to spam e-mails in which machine learning techniques are useful to detect spam e-mails. In another paper, the authors [6] used the text semantic analysis method to classify spam and non-spam e-mails. The proposed method achieved better accuracy as compared to other methods. In this paper, the authors [7] used the remove-replace feature selection technique (RRFST) to remove the features from dataset and achieved better accuracy with the proposed algorithm compared to others. The authors [8] used a novel spam-filtering technique which was based on analyzing the e-mail headers. They used the Hidden Markov Models (HMMs) for analyzing the header structure of e-mails and create a spam detection system. The authors [9] proposed the ALO-Boosting method to classify spam e-mails. In this method, ALO was used for finding the optimum feature subset which gave to boosting algorithm to help for better classification. According to the authors [10], irrelevant messages played a significant role in digital investigations. This message provides a lot of important information for spam e-mails' digital investigation. The authors [11] proposed an efficient algorithm to detect the threads and spam e-mails using the text analytics methodology with the help of e-mail spam corpus. It used the text keyword matching technique with the corpus to classify the spam and it prevents irrelevant mails in the inbox. This paper [12] used the artificial bee colony algorithm with a logistic regression classification model to identify spam e-mails. They used three different publicly available datasets to check their model ability and also compared the performance which is better than the available models. They used feature selection and wrapped methods to develop the technique. The authors [4] proposed the optimization technique to detect spam e-mails. They used the K-nearest neighbors algorithm with distance matrix Euclidean, Manhattan, and Chebyshev to classify the spam e-mails, then used different bio-inspired optimization techniques to classify the spam e-mails and achieved better accuracy. The authors [4] proposed a novel approach to classify the spam e-mails had three steps. In the first step, they used TFDCR FST, the second step described an incremental dynamic model to classify the dataset, and the last third step used the heuristic function to provide a strong ability to recognize the coming e-mails. The authors [13] used a bi-language e-mail text dataset to classify and create a cluster. NGram technique used and achieved better classification results. This paper [14] proposed the GA and RWN technique to detect spam e-mails and also implanted automatic feature detection techniques for better classification. The authors [15] proposed the review paper and read many research papers carefully and extract seven search strategies and got important conclusions related to approaches and techniques. In this paper, the author [16] worked with two real-life datasets to detect opinion spam with the help of a complex probabilistic graph classification approach. They used the neural network technique along with a heterogeneous graph to connect the nodes for concluding about the opinion spam. The main theme of this paper [17] is to classify spam and phishing e-mails. They used the body structure of the e-mail and applies deep-learning and FSTs to identify the e-mails into different categories. The authors [18] used both a supervised regular expression pattern matching technique and an unsupervised K-mean machine learning algorithm to identify the insider threads using analysis of the structure of the e-mails. The author [19] used ID3 and Hidden Markov model to identify spam e-mails and achieved better performance to detect spam e-mails.

This section discussed about SMS classification and twitter spam classification. The authors [20] used two experiments for identifying SMS spam. In the first experiment, they used the Bayesian network classifier method along with the cost-sensitive technique, and in the second experiment, they compared the performance of the proposed technique with existing techniques. The author [21] proposed the MWOA-SPD hybrid technique to detect and classify the tweeter spams.

This section explored the classification of ham and spam documents with Enron dataset. The authors [22] observed security threads about big data in studies. They used the Enron dataset (contains spam and ham documents) to collect a conclusion related to the security issue of big data and also studied in students how they react to spam e-mails. In this research [23], the author preferred the Enron people assignment dataset to create the model and provided the proper recipient. In this paper [24], the author generated a synthetic dataset with the help of the Enron e-mail dataset along with the STDG simulator application to detect threads and knowledge discoveries.

3. PROPOSED FRAMEWORK AND METHODOLOGY

This section discussed the architecture of the proposed framework and methodology. Figure 1 shows the proposed architecture for the classification of spam e-mail. In this architecture, the Enron datasets are collected from the Kaggle repository, and then applied different preprocessing techniques for smoothing the datasets; hence model can achieve better accuracy. The partition of datasets into training and testing is one of the essential steps of the data classification process. This research work has used 10-fold cross-validation for the partition of a dataset into training and testing. We have applied the training and testing of spam and ham messages into different classifiers like Naïve Bayes, J48, RF, Random Tree, and AdaBoosting and evaluate the performance of classifiers in terms of accuracy where RF achieved better accuracy with all seven types of Enron datasets. The RF gained the highest as 98.68% of accuracy with the combined Enron dataset. Finally, we have applied the Chi-Square FST on the combined Enron dataset and reduce the features of datasets. The recommended RF gives the highest as 98.74% of accuracy with reduced features of the combined Enron dataset.

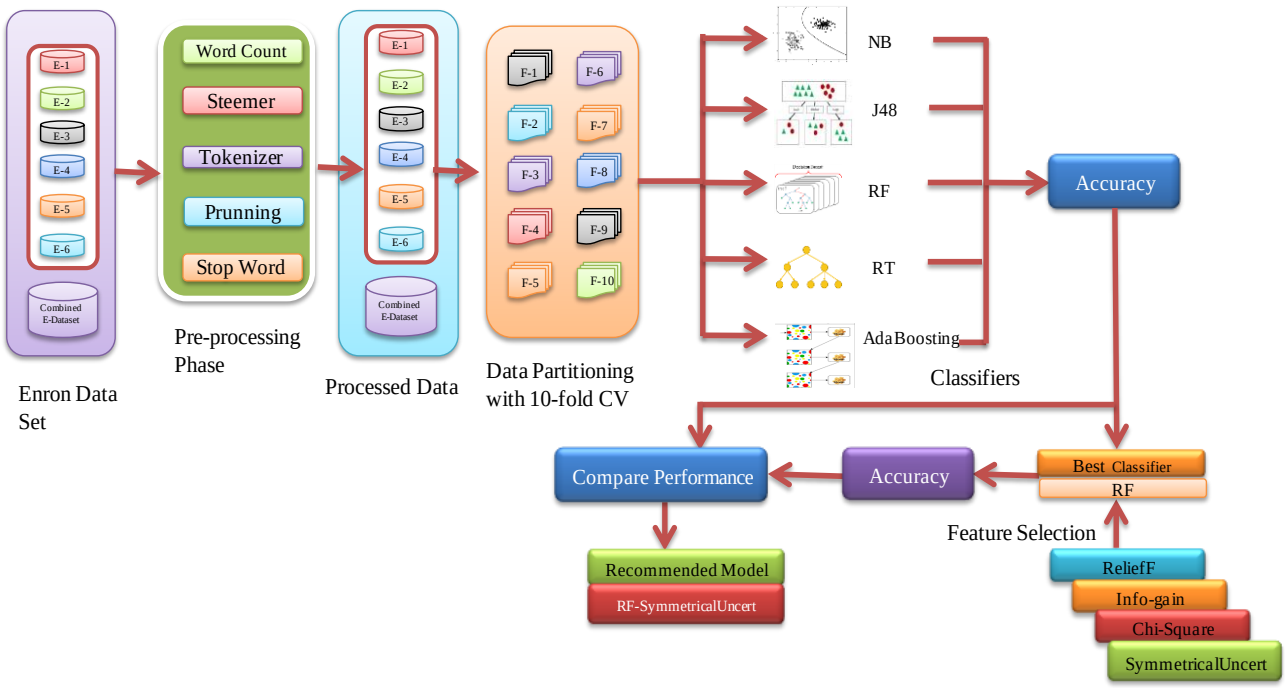

Figure 1. Framework for spam e-mail documents classification

The pseudo code of proposed framework as given below:

Start

Input:

Enron Dataset (Enron1, Enron2, Enron3, Enron4, Enron5, Enron6, Combined Enron Dataset).

The dataset contains spam and ham e-mails documents.

Output:

Performance Measures=PM = {Acc, Sen, Spc, Pr, Fs}

Mbest = Best model, ReducedF_Enron = Enron dataset with reduced features.

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

1. Read Enron datasets
- For input i = 1 to total number of datasets:
- Read Enron(i)
- For input j = 1 to length of datasets:
- do word count and remove stop words, Stemming
- words, tokens and prune dataset
- Record words along with their frequency
- End of inner For
- Return Preprocessed_Enron(i)
- End For

2. Read Preprocessed_Enron datasets
- For input k = 1 to total number of datasets:
- M(k)= Preprocessed_Enron(k) ← {Naïve Bayes, J48, RF, RT,AdaBoosting}
- Mbest(k)=Compare{ (M(k)←Ac) }
- End For
- Mbest→(RF)={Ac, Sen, Spec, Pr, Fs}

3. Read Mbest→(RF)
- For input l = 1 to length of datasets:educedF_Enron = {Preprocessed_Enron}← {Chi-Square, Infogain, ReliefF, SymmetricalUncert}
- Record words along with their frequency
- M(l)= ReducedF_Enron← {RF}
- Mbest(l)=Compare{M(1)←Ac}
- End For
- Mbest(l)={Ac, Sen, Spec, Pr, Fs}

End.

3.1 Enron dataset

Dataset is a significant input for any research. In this research, we have used the Enron dataset collected from the UCI repository. We have collected six different types of Enron datasets. The last seven Enron dataset named as combined Enron dataset is prepared with a combination of collecting all six Enron datasets. Each Enron dataset contains set of the e-mails including spam and ham e-mails documents generated by employees of the Enron Corporation. The main reason for selecting Enron dataset is to develop the robust model that is able to classify the ham and spam documents with high accuracy. The total instances present in each Enron dataset has displayed in a Table 1.

Table 1. Description of Enron datasets

Datasets	Ham-instance	Spam-instance	Total instances
Enron-1	3671	1500	5172
Enron-2	4361	1496	5857
Enron-3	4012	1500	5512
Enron-4	1500	4499	5999
Enron-5	1500	3675	5175
Enron-6	1500	4500	6000
Combine Enron	16383	16383	32766

3.2 Preprocessing

Preprocessing is a very crucial step for the text mining experiment. The big challenge about text mining is to convert unstructured data into structured data. There are several steps involved in transforming unstructured data into structured data.

We follow a number of steps like word count, stemming, tokenization, pruning, and stop word. Word counts measures the total number of words presents in the documents. Stemming means finding the root of words. Tokenization is the process of turning a meaningful piece of data. Pruning is a way to remove unimportant words to improve the structure data and finally use stop words from removing commonly used words like articles (a, an, the, is, etc.), preposition, etc.

3.3 Data partition

The data partition is a technique to partition the data into training and testing sets. Cross-validation [25] is a data partition technique used to analyze the performance of the machine learning techniques. We have used a 10-fold cross-validation technique to create a partition of the dataset into ten equal parts, randomly selecting one part of the dataset as a testing dataset. The remaining 9 part is for the training dataset. This iteration occurs ten times to perform the analysis of the algorithm.

3.4 Classification techniques

There are different classification algorithm resides in the WEKA tools (http:// www.cs.waikato.ac.nz/~ml/weka/). This algorithm provides various principles to classify the text data. There are different algorithm likes Naive Bayes, J48, RF, Random Tree, and Adaboosting.

3.4.1 Naïve Bayes

Naïve Bayes [26] classifier strictly follows the Bayes' theorem principle. Bayes theorem works the conditional probability principle. It used posterior probability in this kind of method. Naive Bayes is useful for the large size of the dataset.

3.4.2 J48

J48 [27] is also known as C4.5. The C4.5 algorithm is the successor of ID3 (iterative Dichotomiser). C4.5 follows the non-backtracking approach of the greedy method in which the decision tree builds via top-down recursive and it also follows the divide-and-conquer methodology. All the tuples are associated with the class labels. We have used the training set to build the tree using the recursively partitioned method.

3.4.3 Random Forest

Random Forest (RF) [28] follows the principle of a supervised machine learning algorithm. As the name suggests, a RF comprises numerous decision trees. The model prediction depends upon the class of the most vote which come from the decision tree, that's why random forests follow the wisdom of crowds. RF is useful for large-size datasets and has a large number of input features. The classification problem can handle by either the Gini-index or the entropy in a RF.

$$\text{Gini} = 1 - \sum_{k=0}^{n} (Xi)^k \quad (1)$$

When we use the entropy to find the decision node in a random tree can be expressed like

$$\text{Entropy} = \sum_{k=0}^{n} -Xi * log2(Xi) \quad (2)$$

Xi represents the relative frequency of the class that observes in the dataset, and n represents the number of classes.

3.4.4 Random tree

The random tree [26, 29] has been introducing by Leo Breiman and Adele Cutler. A random tree algorithm is helpful to solve both regression and classification problems. The random tree belongs to the supervised learning group. It is an ensemble learning algorithm that generates many individual learners. It enlists a bagging scheme to generate a random set of data for building a decision tree.

3.4.5 Adaboosting

The abbreviation of Adaptive Boosting [30] is AdaBoost. The AdaBoost algorithm was the first algorithm created for binary classification. With the help of the AdaBoost algorithm easily increase the performance of any machine learning algorithm. It is very helpful for weak learners. The model which used the AdaBoost algorithm attains better accuracy as compared to other classification problems. The equation is represented by

$$\text{F(x)} = Sign\left(\sum_{k=1}^{K} Wk * Fk(x)\right) \quad (3)$$

where, Fk represents the Kth weak classifier and Wk represents the corresponding weight.

3.5 Feature selection

Feature selection is a technique that selects the most valuable feature among all features. In this technique, we select all the features that have a significant effect and remove all the features that do not have a significant effect. In this research, we use the FST followed by the ranker method.

Chi-square [31] used the relation between feature and category of words. The feature means the occurrence of frequency of feature and category means the probability of occurrence of category. We find the correlation between feature and category, if they are dependent then find the level of correlation between them, and if independent then we do not apply the FST. SymmetricalUncert [32] FST evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class. This method follows the ranking technique to calculate the rank of the feature. It can remove the feature whose value is less than the threshold value. Information Gain (IG) [33] is an entropy-based FST, followed by the ranker method. IG is defined as the amount of information provided by the feature items for the text category. IG is calculated by how much of the term can be used for the classification of information, in order to measure the importance of lexical items for the classification. ReliefF [34] FST was the extension of the Relief feature selection algorithm in 1994. ReliefF can handle the classification of multi-class data. ReliefF selects the random sample from the training set. After that, it finds out k near hits from the same class and k near misses from each different class. In this way, it updates the weight of the feature and repeats this process n times to increase the weights of all features.

4. EXPERIMENTAL RESULTS

In this section, we present the results related to our experiment. We accomplished the experimental work using the WEKA tool in Windows 10 operating system environment. In this research work, we have used six Enron datasets to find the classification accuracy between ham and spam e-mails.

This experiment is divided into two parts: In the first part, we take six different Enron datasets, and the second part combines all six datasets and creates one signal combined dataset. Table 2 shows the number of instances of Enron datasets. We have applied pre-processing techniques like word count, steamer, tokenizer, pruning, and stop word to remove the unimportant words from both kinds of datasets as shown in Table 3. Table 3 shows the remaining words after applying preprocessing technique in datasets. Figure 2 is the graphical representation of the words related to all datasets. After that, we apply machine learning algorithms like Naïve Bayes, J48, RF, Random Tree, and Adaboosting to get accuracy which is shown in Table 4. Table 4 shows that the accuracy of classifiers like Naïve Bayes, J48, RF, Random Tree, and Adaboosting with each dataset and also get one more conclusion that RF achieves the highest accuracy with each dataset. Figure 3 is the graphical representation of accuracy related to each dataset. In Table 5, we represent the confusion matrix related to the highest accuracy achieved by the RF algorithm with the combined Enron dataset. We have also calculated different performance measures like true positive rate (TPR), false positive rate (FPR), precision, specificity, and F-measures, which are represented in Table 6. In the next section, we apply the FSTs like ReliefF, Info-gain, Chi-square, and SymmetricalUncert followed by the ranker method in the combined Enron dataset with RF machine learning algorithm using a 10 fold cross-validation data partition method and achieved better accuracy with 10000 features. The RF achieved an accuracy of 98.73% is the highest among other accuracy with SymmetricalUncert FSTs in the case of the combined Enron dataset as shown in Table 7. We show the confusion matrix and different performance measures of RF in Table 8 and Table 9 respectively. Figure 4 shows that the performance measures of the best classifier RF with different FSTs.

The above Table 3 and Figure 2 show that the relevant numbers of words in datasets after applying preprocessing techniques like wordcounts, steamer, tokenizer, pruning and stop word.

Table 4 and Figure 3 show that the accuracy of different classifiers with different Enron datasets where Random Forest classifiers gives best accuracy compared to others.

Table 9 and Figure 4 show that the various performance measures of Random Forest classifiers with different FSTs. Finally, this research work has compared the performance of our suggested RF classifier with SymmetricalUncert FST to the existing models where suggested RF achieved better accuracy compared to others. In the base paper [3], the author used a combination of six Enron datasets and got 98.57% of accuracy but our suggested classifier RF achieved 98.68% of accuracy, and at the same time, we have applied SymmetricalUncert FST on a combined Enron dataset where RF got 98.73% of accuracy with reduced feature subsets.

Table 2. Number of instances of Enron datasets

Name of datasets	Number of instances
Enron-1	5172
Enron-2	5857
Enron-3	5512
Enron-4	5999
Enron-5	5175
Enron-6	6000
Combined Enron	32766

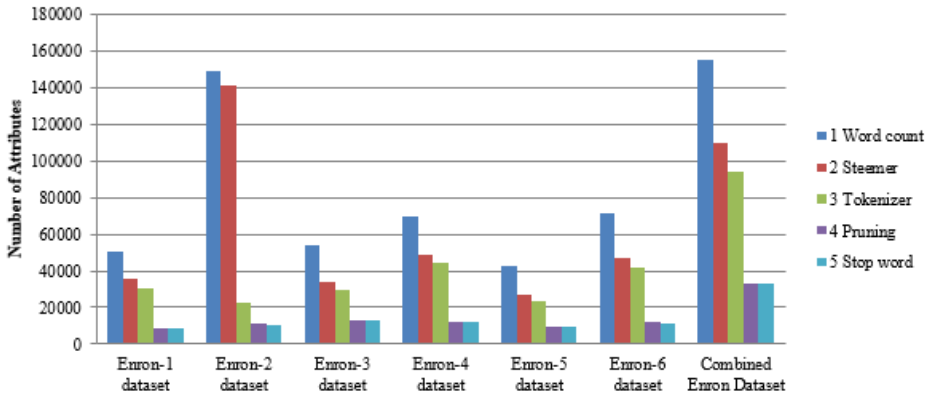

Figure 2. Graphical representation of preprocessed datasets

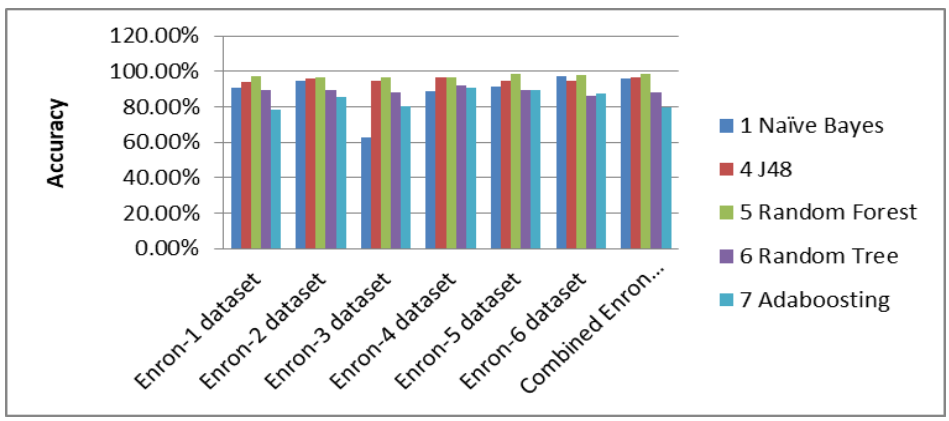

Figure 3. Graphical representation of classifiers accuracy with different Enron datasets

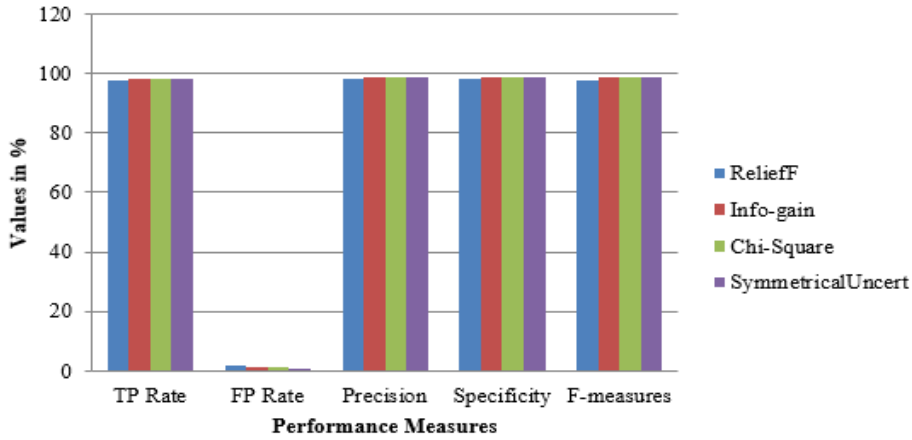

Figure 4. Performance measures of best classifier RF with FST in case of combined Enron dataset

Table 3. Preprocessing of Enron datasets

Sr. No.	Preprocessing steps	Number of attributes						
		Enron-1 dataset	Enron-2 dataset	Enron-3 dataset	Enron-4 dataset	Enron-5 dataset	Enron-6 dataset	Combined Enron Dataset
1.	Word count	50557	148914	53872	69528	42285	71597	155115
2.	Stemmer	35623	141393	33864	48969	27335	47158	109459
3.	Tokenizer	30852	22214	29726	44176	23694	41838	94498
4.	Pruning	8763	10850	13030	12219	9333	11811	33517
5.	Stop word	8529	10612	12788	11980	9100	11575	33264

Table 4. Accuracy of the classifiers with Enron datasets

Sr. No.	Classifier	Enron-1 dataset	Enron-2 dataset	Enron-3 dataset	Enron-4 dataset	Enron-5 dataset	Enron-6 dataset	Combined Enron Dataset
1	Naïve Bayes	90.93 %	94.83 %	62.94 %	88.78 %	91.49 %	97.12 %	96.15 %
2	J48	93.72 %	95.99 %	94.92 %	96.38 %	94.38 %	94.43 %	96.38%
3	Random Forest (RF)	**97.53 %**	**96.38 %**	**96.83 %**	**96.48 %**	**98.59 %**	**97.97 %**	**98.68 %**
4	Random Tree	89.23 %	89.57 %	88.28 %	92.12 %	89.60 %	86.45 %	87.92 %
5	Adaboosting	78.65 %	85.66 %	80.08 %	90.55 %	89.31%	87.55 %	79.89 %

Table 5. Confusion matrix of best RF classifier with Enron datasets

Actual vs. Predicted	Enron-1 dataset		Enron-2 dataset		Enron-3 dataset		Enron-4 dataset		Enron-5 dataset		Enron-6 dataset		Combine Enron dataset	
	Ham	Spam	Ham	Spam	Ham	Spam	Ham	Spam	Ham	Spam	Ham	Spam	Ham	Spam
Ham	3602	70	4347	14	4008	4	1289	211	1429	71	1379	121	16143	240
Spam	58	1442	198	1298	171	1329	0	4499	2	3673	1	4499	194	16189

Table 6. Performance measures of best RF classifier with different Enron datasets

Name of Measures	Datasets						
	Enron-1	Enron-2	Enron-3	Enron-4	Enron-5	Enron-6	Combined Enron
TP Rate	98.09	99.68	99.90	85.93	95.27	91.93	98.54
FP Rate	3.87	13.24	11.40	0	0.05	0.02	1.18
Precision	98.42	95.64	95.91	100	99.86	99.93	98.81
Specificity	96.13	86.76	88.6	100	99.95	99.98	98.82
F-measures	98.25	97.62	97.86	92.43	97.51	95.76	98.67

Table 7. Accuracy of RF with FSTs with Combined Enron datasets

Feature Selection Techniques	Accuracy in %
ReliefF	97.88
Info-gain	98.68
Chi-Square	98.71
SymmetricalUncert	98.73

Table 8. Confusion matrix RF with FSTs in case of Combined Enron datasets

Actual	ReliefF		Info-gain		Chi-Square		SymmetricalUncert	
Vs. Predicted	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**	**Spam**
Ham	15976	407	16118	265	16126	257	16122	261
Spam	288	16095	167	16216	165	16218	156	16227

Table 9. Performance measures of RF with FSTs in case of Combined Enron datasets

Name of attributes	**ReliefF**	**Info-gain**	**Chi-Square**	**SymmetricalUncert**
TP Rate	97.52	98.38	98.43	98.41
FP Rate	1.76	1.02	1.01	0.95
Precision	98.23	98.97	98.99	99.04
Specificity	98.24	98.98	98.99	99.05
F-measures	97.87	98.67	98.71	98.72

5. CONCLUSION AND FEATURE WORKS

Classification of spam e-mail documents with high accuracy is a very challenging task for researchers. In this paper, we have prepared a model related to a spam e-mails documents classification. We have used seven Enron datasets, which are Enron-1, Enron-2, Enron-3, Enron-4, Enron-5, Enron-6, and combined Enron datasets. This research work has used classification techniques to analyze the Enron dataset and classify the spam and ham e-mails documents. We have used Naïve Bayes, J48, RF, Random Tree, and AdaBoosting algorithms to develop a model. The RF classifier achieves the highest accuracy in all seven datasets compared to other algorithms. The RF classifier gives the highest 98.68% of accuracy in the case of the combined Enron dataset. Then, we apply the SymmetricalUncert FST on the combined Enron dataset and classify and analyzed it with a RF algorithm where RF achieves better accuracy as 98.73%. Finally, we have suggested that RF with the SymmetricalUncert FST model is better for the classification of spam e-mails documents. In the future, we will develop a robust and computationally intelligent hybrid model which will give better accuracy compared to others in the field of spam filtering, phishing e-mails classification and different types of attacks. We will also develop new feature selection and optimization techniques which will reduce the irrelevant number of features (words) from dataset and achieve the better classification accuracy with a smaller number of features and less computational time.

ACKNOWLEDGMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that help improve the quality of this work.

REFERENCES

[1] Weblink: https://en.wikipedia.org/wiki/History_of_email_spam, accessed on 22 July 2021.

[2] Kaspersky Lab Spam Report. (2021). https://securelist.com/spam-and-phishing-in-q2-2021/103548/, accessed on 20 Aug. 2021.

[3] Saleh, A.J., Karim, A., Shanmugam, B., Azam, S., Kannoorpatti, K., Jonkman, M., Boer, F.D. (2019). An intelligent spam detection model based on artificial immune system. Information.s Information, 10(6): 1-17. https://doi.org/10.3390/info10060209

[4] Batra, J., Jain, R., Tikkiwal, V.A., Chakraborty, A. (2021). A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. Int. J. Inf. Manag. Data Insights, 1(1): 1-13. https://doi.org/10.1016/j.jjimei.2020.100006

[5] Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid,S.M., Adetunmbi, A.O., Ajibuwa, O.E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. Heliyon, 5(6): 1-23. https://doi.org/10.1016/j.heliyon.2019.e01802

[6] Saidani, N., Adi, K., Allili, M.S. (2020). A semantic-based classification approach for an enhanced spam detection. Comput. Secur., 94: 1-12. https://doi.org/10.1016/j.cose.2020.101716

[7] Hota, H.S., Shrivas, A.K., Hota, R. (2018). An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique. Procedia Comput. Sci., 132: 900-907. https://doi.org/10.1016/j.procs.2018.05.103

[8] Salcedo-Campos, F., Díaz-Verdejo, J., García-Teodoro, P. (2012). Segmental parameterisation and statistical modelling of e-mail headers for spam detection. Inf. Sci. (Ny)., 195: 45-61. https://doi.org/10.1016/j.ins.2012.01.022

[9] Naem, A.A., Ghali, N.I., Saleh, A.A. (2018). Antlion optimization and boosting classifier for spam email detection. Futur. Comput. Informatics J., 3(2): 436-442. https://doi.org/10.1016/j.fcij.2018.11.006

[10] Yu, S. (2015). Covert communication by means of email spam: A challenge for digital investigation. Digit. Investig., 13: 72-79. https://doi.org/10.1016/j.diin.2015.04.003

[11] Murugavel, U., Santhi, R. (2020). Detection of spam and threads identification in E-mail spam corpus using content based text analytics method. Materials Today: Proceedings, 33: 3319-3323. https://doi.org/10.1016/j.matpr.2020.04.742

[12] Dedeturk, B.K., Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Appl. Soft Comput. J., 91: 106229. https://doi.org/10.1016/j.asoc.2020.106229

[13] Alsmadi, I., Alhami, I. (2015). Clustering and classification of email contents. J. King Saud Univ. - Comput. Inf. Sci., 27(1): 46-57.

https://doi.org/10.1016/j.jksuci.2014.03.014
[14] Faris, H. (2019). An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks. Inf. Fusion, 48: 67-83. https://doi.org/10.1016/j.inffus.2018.08.002
[15] Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., Odusami, M. (2019). A review of soft techniques for SMS spam classification: Methods, approaches and applications. Eng. Appl. Artif. Intell., 86: 197-212. https://doi.org/10.1016/j.engappai.2019.08.024
[16] Liu, Y., Pang, B., Wang, X. (2019). Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph. Neurocomputing, 366: 276-283. https://doi.org/10.1016/j.neucom.2019.08.013
[17] Rastenis, J., Ramanauskaitė, S., Suzdalev, I., Tunaitytė, K., Janulevičius, J., Čenys, A. (2021). Multi-language spam/phishing classification by email body text: Toward automated security incident investigation. Electron., 10(6): 1-10. https://doi.org/10.3390/electronics10060668
[18] Michael, A., Eloff, J.H.P. (2019). A machine learning approach to detect insider threats in emails caused by human behaviours. Proceedings of the Thirteenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2019), 34-49.
[19] Kumar, V., Monika, Kumar, P., Sharma, A. (2018). Spam email detection using ID3 algorithm and hidden Markov Model. 2018 Conf. Inf. Commun. Technol. CICT 2018, pp. 1-6. https://doi.org/10.1109/INFOCOMTECH.2018.8722378
[20] Lim, L.P., Singh, M. (2020). Resolving the imbalance issue in short messaging service spam dataset using cost-sensitive techniques. J. Inf. Secur. Appl., 54: 1-10. https://doi.org/10.1016/j.jisa.2020.102558
[21] Krithiga, R., Ilavarasan, E. (2020). A reliable modified whale optimization algorithm based approach for feature selection to classify twitter spam profiles. Microprocess. Microsyst. https://doi.org/10.1016/j.micpro.2020.103451
[22] Zaki, T., Uddin, M.S., Hasan, M.M., Islam, M.N. (2017). Security threats for Big Data. Int. Conf. Res. Innov. Inf. Syst. ICRIIS, pp. 1-6. https://doi.org/10.1109/ICRIIS.2017.8002481
[23] Rameshkumar, R., Bailey, P., Jha, A., Quirk, C., (2019). Assigning people to tasks identified in email: The EPA dataset for addressee tagging for detected task intent. Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, Brussels, Belgium, pp. 28-32. https://doi.org/10.18653/v1/w18-6104
[24] Babalola, K.O., Jennings, O.B., Urdiales, E., Debardelaben, J.A. (2019). Statistical methods for generating synthetic email data sets. Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018, 3986-3990. https://doi.org/10.1109/BigData.2018.8622601
[25] Laorden, C., Santos, I., Sanz, B., Alvarez, G., Bringas, P.G. (2012). Word sense disambiguation for spam filtering. Electron. Commer. Res. Appl., 11(3): 290-298. https://doi.org/10.1016/j.elerap.2011.11.004
[26] Gupta, A., Mohan, K.M., Shidnal, S. (2018). Spam filter using naïve bayesian technique. Int. J. Comput. Eng. Res., 8(6): 2250-3005.
[27] Yadav, A.K., Chandel, S.S. (2015). Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model. Renew. Energy, 75: 675-693. https://doi.org/10.1016/j.renene.2014.10.046
[28] Aulia, A., Jeong, D., Saaid, I.M., Kania, D., Shuker, M.T., El-Khatib, N.A. (2019). A random forests-based sensitivity analysis framework for assisted history matching. J. Pet. Sci. Eng., 181: 106237. https://doi.org/10.1016/j.petrol.2019.106237
[29] Kalmegh, S. (2015). Analysis of WEKA data mining algorithm reptree, simple cart and randomtree for classification of Indian News. Int. J. Innov. Sci. Eng. Technol., 2(2): 438-446.
[30] Mazini, M., Shirazi, B., Mahdavi, I. (2019). Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms. J. King Saud Univ. - Comput. Inf. Sci., 31(4): 541-553. https://doi.org/10.1016/j.jksuci.2018.03.011
[31] Rachburee, N., Punlumjeak, W. (2015). A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. Proc. - 2015 7th Int. Conf. Inf. Technol. Electr. Eng. Envisioning Trend Comput. Inf. Eng. ICITEE 2015, 420-424. https://doi.org/10.1109/ICITEED.2015.7408983
[32] Dai, J., Chen, J., Liu, Y., Hu, H. (2020). Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation. Knowledge-Based Syst., 207: 106342. https://doi.org/10.1016/j.knosys.2020.106342
[33] Lei, S. (2012). A feature selection method based on information gain and genetic algorithm. Proc. - 2012 Int. Conf. Comput. Sci. Electron. Eng. ICCSEE 2012, 2: 355-358. https://doi.org/10.1109/ICCSEE.2012.97
[34] Yang, F., Cheng, W., Dou, R., Zhou, N. (2011). An improved feature selection approach based on ReliefF and mutual information. 2011 Int. Conf. Inf. Sci. Technol. ICIST 2011, pp. 246-250. https://doi.org/10.1109/ICIST.2011.5765246

NOMENCLATURE

FST	Feature Selection Technique
RF	Random Forest
TP	True Positive
FP	False Positive