International Journal of Management, IT & Engineering
Vol. 11 Issue 09, September 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Ensemble Classification Approach for Email Spam Filtering

Shivani Thakur
Department of CSE
IEC University
Baddi, India

Randeep Singh
Department of CSE
IEC University
Baddi,India

Ravinder Madhan
Department of CSE
IEC University
Baddi, India

Abstract- **E-mail has become a significant mean of message sharing. A perfect way is obtained through email for sending this enormous volume of ads without any expenditure for the forwarder and bitter reality is a number of organizations make the utilization of this in these days. The email spam detection has various phases which include pre-processing, feature extraction and classification. In this voting classification is proposed which is the combination of random forest, naive bayes and SVM. The proposed method is implemented in python and results are analyzed in terms of accuracy, precision and recall**.

Keywords-Voting Classification, Random Forest, Naive Bayes, SVM, Email Spam

I. INTRODUCTION

Data mining, an important step in KDD (knowledge discovery in databases) is considered to be the process concerned with the analysis of enormous data stored in the data warehouses, and the discovery of the inherent information of huge potential [1]. Data mining, by digging large amounts of data, can potentially discover concealed relationships and disclose unknown patterns and trends. Based on the performed work, the tasks of data mining, or models, are generally categorized into four categories of association, classification, clustering, and regression.Analysisof Data mining typically relies upon three techniques: classical statistics, artificial intelligence, and machine learning [2]. Classical statistics is largely used to analyze data, data relations, and to deal with numerical data in big databases. Some popular classical statistics approaches are regression analysis, clustering, and discriminate analysis. Presently, E-mail has becomea significant mean of message sharing. A perfect way is obtained through email for sending this enormous volume of ads without any expenditure for the forwarder and bitter reality is a number of organizations make the utilization of this in these days[3][4]. Consequently, these unwanted bulk e-mails considered as a spam or junk mail create confusion in the e-mailboxes of millions of people. Spam is incredibly inexpensive to send. Thus, much problem is occurred before the Internet community. The delay is occurred while delivering the authentic email due to the enormous spam-traffic among the servers. People having dial-up Internet access need to use bandwidth downloading garbage e message. Much time is consumed in the separation of unwanted messages and a risk is put forward of erasing ordinary mail accidentally. At last, a measure of explicit spam is also there that must not be exposed to children [5].

Unluckily, there is not any universal and perfect scheme to remove the spam exists due to which amount of junk mail is maximized rapidly. To illustration, half of messages received in the personal mailbox are spam[6]. The mail can be filtered using 2 general approaches such as KE and ML. In the previous condition, a bunch of guidelines is generated on the basis of which the classification of messages is done as spam or legitimate mail. The major shortcoming of this technique is the requirement of constantly update and its maintenance is not convenient for most of the users[7]. The maintainer of the spam filtering tool is allowed to update the rules in a centralized way. A P2P knowledgebase solution is also available. In case of presence of rules publically, the spammer is capable of adjusting the text of that message to let it pass

through the filter[8]. Thus, customization of spam filtering is superior on a per-user basis. There is not any need to specify the rules in explicit manner using ML. Despite, a bunch of already classified records is required. Afterward, a particular algorithmic strategy is implemented for learning the classification rules from this data. The study of subject of ML is carried out extensively and various algorithms can be utilized for this task [9].The broadcast of spam whether or not its payload is delivered and applied, has many negative impacts. Some of these impacts have been discussed below:

a) Direct Impacts: Spam offers an uncontrolled communication channel, using which spammers can deceive targets, sell substandard goods, inject viruses, etc. These impacts are large scale, but are not specifically allowed by the victims. For example, the computer of the victim can be used in further spamming or to trigger a cyber-attack. In the same way, the stolen identity of the victim can be used in criminal activity, and against other targets [10].

b) Network Resource Consumption: The bulk of email traffic is spam today. This consumption of bandwidth and storage by this traffic is very high, which, in turn, increases the risk of untimely delivery of messages or absolute message loss.

c) Human Resource Consumption: Sorting a spam-filled inbox is an unfriendly experience, and needs considerable time too. This process certainly restricts the timeliness of the email as the recipient would otherwise be sorting through spam. In addition, frequent arrival of spam can prevent the use of email arrival alerts, enforcing a rule of batch rather than reading incoming emails, compromising further timeliness [11].

d) Lost Email: Valid email may be lost or ignored due to spam. Certainly, spam cancellation methods may be the reason of the vanish of legitimate emails. Generally, spam scorns the use of email and consequentlyusers get discourage of using it. Users can be reluctant to split their email addresses or interrupt them in means that prevent the use of email as a channel to interact them [12].

II. Literature Survey

Da Xiao, et.al (2020) designed a secure mail system in which KNN (k-nearest neighbour) algorithm and improved LSTM (long short-term memory) algorithm recognized as Bi-LSTM-Attention algorithm were implemented [13]. The normal emails were differentiated from the spam and phishing emails in effective manner using KNN and provided higher accuracy. The phishing emails were classified using Bi-LSTM-Attention model on the basis of the similarity of the malicious mail text taken from the same attacker to some extent. The source of malicious emails were classified and recognized to grasp the features of the attacker, providing the materials for further research and enhancing the passive status of users. The experimental results exhibited that the designed system provided 90% accuracy.

Fahd Aldosari, et.al (2018) recommended a technique on the basis of nonparametric Bayesian inference for which an infinite scaled Dirichlet mixture model was deployed [14]. The scaled Dirichlet was taken as a flexible generalization of the popular Dirichlet distribution. The Markov Chain Monte Carlo and a variation nal Bayes model were applied in order to learn the resulting model in this technique. Different from the earlier techniques, the recommended technique considered the visual content that was ignored instead it was utilized through the spammers. The results obtained in experiments exhibited that the recommended technique was beneficial for the email spam filtering.

Bilge KaganDedeturk, et.al (2020) presented a new spam detection technique in which the ABC (artificial bee colony) algorithm was put together with a LR (logistic regression) classifier [15]. The results generated on public datasets validated that the presented technique was capable of handling the

high-dimensional data and represented the local and global search potentials of this technique. The efficiency of this technique to detect the spam was compared with SVM, LR and NB models. It was observed that the presented technique performed more effectively with regard to the accuracy.

Samira Douzi, et.al (2017) intended a novel Spam filter on the basis of PV-DM (Paragraph Vector-Distributed Memory) for dealing with the drawbacks of the BoW (Bags of Words) representation [16]. A system was suggested in which the representation of context of an email was incorporated with its selected attributes. A more comprehensive filter was presented in this intended filter in order to classify the emails and this approach was implemented on the python programming language. The public datasets named EM Canada18 or GenSpam19 were applied to deploy this approach. The experimental outcomes revealed that the intended approach provided better outcomes.

El-Sayed M. El-Alfy, et.al (2016) projected an intelligent model in order to filter the multimodal textual communication that had contained emails and short messages [17]. A new system that had influence of human immune system and hybrid techniques of ML was exploited for fusing the information. The relevant attributes were extracted and chosen using various techniques so that the complexity of the projected model was mitigated for suiting the mobile applications when the good performance was preserved. Different datasets were utilized to compute the projected model on the basis of results.

VitorBasto-Fernandes, et.al (2016) constructed an approach in which the traditional algorithm was extended in the domain of spam filtering [18]. For this, three new indicated based SMS-EMOA algorithm, CH-EMOA and decomposition-based MOEA/D evolutionary multi-objective algorithm were deployed. The performance of a heterogeneous ensemble of classification algorithms was optimized into two different but complementary scenarios such as parsimony maximization and e-mail classification under low confidence level. The experimental outcomes acquired on the public standard corpus depicted that the constructed approach was suitable in the spam filtering domain.

Maria Habib, et.al (2018) suggested an efficient email spam detection framework on the basis of GP (Genetic Programming) along with SMOTE (Synthetic Minority Over-sampling Technique) for detecting the spam emails [19]. The extraction of attributes of datasets was done from the public spam corpora. Two benchmark email corpora were applied to implement and test this approach and 4 other well-recognized classification algorithms were employed for testing this approach with regard to accuracy, recall, precision and G-mean. The results of experiments demonstrated that the suggested approach was more efficient for classifying the spam emails as compared to the usual classifiers. The future work would focus on expanding this work and analyzing the relative importance of the attributes.

KritiAgarwal, et.al (2018) introduced an integrated approach of ML (machine learning) named NB (Naive Bayes) algorithm and computational intelligence based PSO (Particle Swarm Optimization) in order to detect the email spam [20]. The email content was learned and classified as spam and non-spam using NB algorithm. The stochastic distribution and swarm behaviour property had contained in the PSO algorithm. Thus, this algorithm was applied to optimize the metrics of Naive Bayes algorithm globally. Ling spam dataset was applied to conduct the experiments and compute the performance of introduced approach concerning precision, recall, f-measure and accuracy. The results proved the superiority of Particle Swarm Optimization algorithm over the individual NB algorithm.

NishthaJatana, et.al (2014) developed an encoded and fragmented database approach which was similar to the radix sort method and implemented to Paul Graham's NB (Naive Bayes) algorithm in order to filter the email spam [21]. This approach emphasized on mitigating the overall time in the process of detecting the spam in email. Two publicly available datasets were executed to analyze the developed

approach quantitatively and qualitatively and showed that the time performance was enhanced. This approach was performed six times faster as compared to traditional algorithm. The outcomes exhibited that the developed approach provided quicker performance than traditional Paul Graham's spam filtering method.

Siti-Hajar-Aminah Ali, et.al (2015) established a RAN-LSH (Resource Allocating Network with Locality Sensitive Hashing) algorithm with data selection [22]. The data was selected by searching a hash table with similar spam emails whose learning was done earlier. For this purpose, Locally Sensitive Hashing was executed. The novel spam emails were discovered using an outlier detection system. The BoW (Bag-of-Words) algorithm was adopted for the analysis of email contents and the feature vectors were produced in which features were converted on the basis of normalized term frequency-inverse document frequency. The established approach was computed on the dataset of double-bounce spam emails. The results exhibited that the established algorithm was efficient to enhance the accuracy over the testing period with the help of outlier detection algorithm.

Vikrant Kumar, et.al (2018) projected the ID3 algorithm to generate the DT (decision tree) and the HMM (Hidden Markov Model) to count the probabilities of the occurrence of the events so that the emails were classified as spam or ham [23]. All the posteriorly classified words of emails were utilized to assign the label of spam or ham to emails. Thereafter, the DTs were generated to supervise all the processed emails. An Enron dataset that contained 5172 emails in which 2086 emails were spam and 2086 were ham had implemented to quantify this algorithm. The experimental outcomes indicated that the projected algorithm provided the accuracy around 89% on the spam emails.

Ahmed I. Taloba, et.al (2019) recommended a hybrid technique GADT in which PCA (principle component analysis) model for removing the irrelevant attributes that had less processing [24]. A hybrid approach that had GA (genetic algorithm) and DT (decision tree) algorithms was exploited for developing this hybrid approach. In this, the performance of standard DT algorithm was enhanced using GA. The PCA algorithm was deployed for alleviating the high dimensionality of the feature vector which assisted in increasing the performance of all the text classification algorithms using which the spam of email was detected. The experimental results validated that the recommended technique was efficiently enhanced the accuracy for detecting the spam of emails in contrast to conventional DT.

III. Research Methodology

E-mail is the most significant mean of communication. A perfect way is obtained through email for sending these enormous ads without any expenditure for the forwarder and bitter reality is a number of organizations make the utilization of this in these days. Consequently, these unwanted bulk e-mails considered as a spam or junk mail create confusion in the e-mailboxes of millions of people. Spam is incredibly inexpensive to send. Thus, much problem is occurred before the Internet community.

Email spam prediction process includes four steps which are:

A. Data Acquiring: To conduct tests, the data is outsourced from different associations in healthcare.

B. Data preprocessing: Enforcing machine learning methodologies to such an extent that conclusion can be presented and a significant examination can be accomplished on the data to preprocess it. This level delivers clean and demised information for the feature selection step by eliminating repetitive traits from the dataset for making the training framework more competent.

C. Feature selection: This step involves using a subset that consists of highly exclusive attributes for predicting spam emails. These specific features belong to current class of features. The introduced methodology selects features by applying the random forest algorithm. This framework accepts 100 as the estimator rate and creates tree design of the most significant attributes. RF classification picks those

International Journal of Management, IT & Engineering
Vol. 11 Issue 09, September 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

attributes which show up generally fitting or important for foreseeing email spam.

D. Classification: To predict spam emails, The picked features are mapped to the training model and further classified. A sort of email spam is addressed by each different class. The logistic regression framework is enforced for classifying features. It takes extracted features as input. In this project, mails are classified into two classes known as spam and ham.

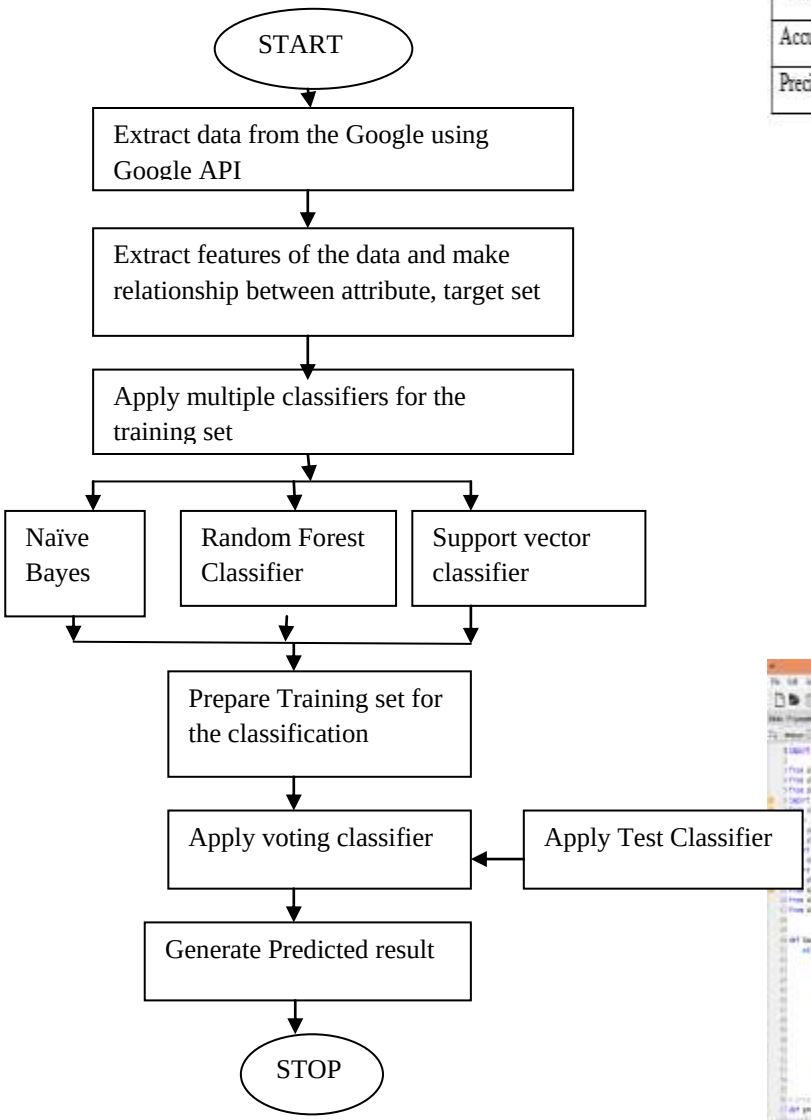

Fig 2: Proposed Flowchart

Result and Discussion

The results will be analyzed on the following performance metrics which are shown in table 4.1 and explained as follows:

Table 1 Performance Parameter

False Positive Rate (FPR)	FPR = FP/(FP+TN)
False Negative Rate (FNR)	FNR = FN/(TP+FN)
True Positive Rate (TPR)	TPR = TP/(TP+TN)
True Negative Rate (TNR)	TNR = TN/(TN+FP)
Accuracy	(TP+TN)/(TP+TN+FN+FP)
Precision	TP/TP+FP

- True Positive (TR) is described as the amount of positive tuples that are accurately labeled by the classifier.
- True Negative (TN) is defined as the amount of negative tuples that are accurately labeled by the classification model.
- False Positive (FP) is defined as the negative tuples that are wrongly labeled as positive.
- False Negative (FN) is defined as the positive tuples that are wrongly labeled as negative.

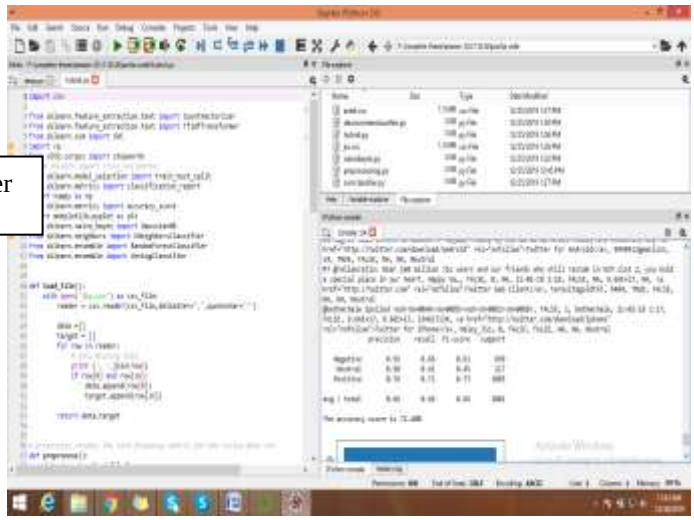

Fig 3 Apply Voting Classifier

International Journal of Management, IT & Engineering
Vol. 11 Issue 09, September 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Fig 3 show that the email spam detection has various steps. The classification technique can classify data into positive, negative and neutral classes. The voting classification model also classifies data into similar classes. The voting classification model classifiers give 72.48 percent accuracy

Table 2 Comparison Analysis

Parameter	Naïve Bayes	SVM	Decision Tree	Voting
Accuracy	47.68 Percent	65.62 Percent	62.12 percent	72.48 percent
Recall	0.48	0.64	0.62	0.66
Precision	0.64	0.65	0.63	0.66

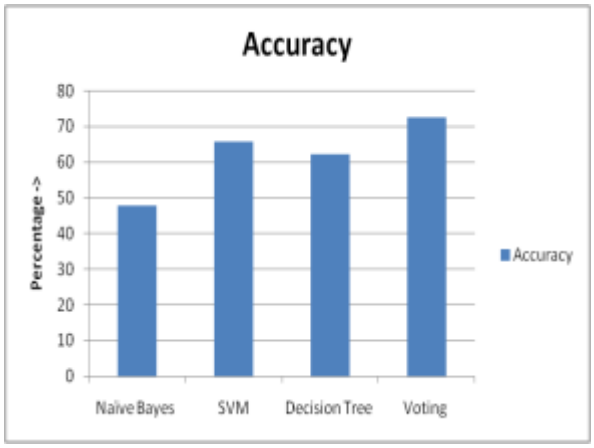

Fig 4 Accuracy Analysis

Fig 4, the accuracy of various classifiers like NB, SVM, Decision and voting classifiers are compared for the email spam detection. The voting classifier has maximum accuracy which is approx 72 percent as compared to other classifiers.

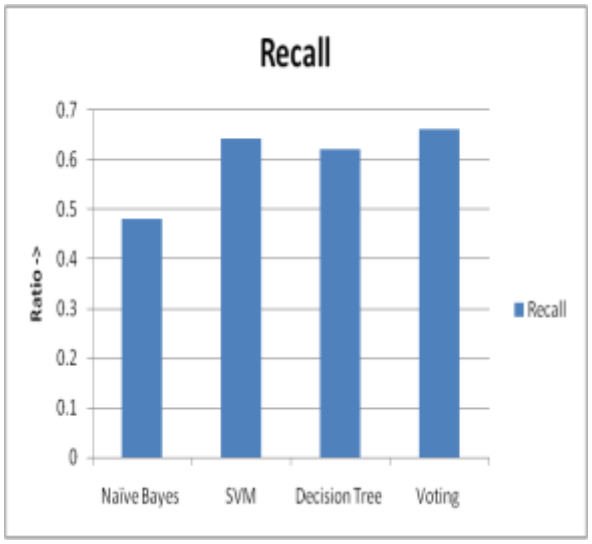

Fig 5 Recall Analysis

As shown in fig 5, the recall of various classifiers is compared for the email spam detection. The recall classifier has maximum recall which is approx 0.67 as compared to other classifiers

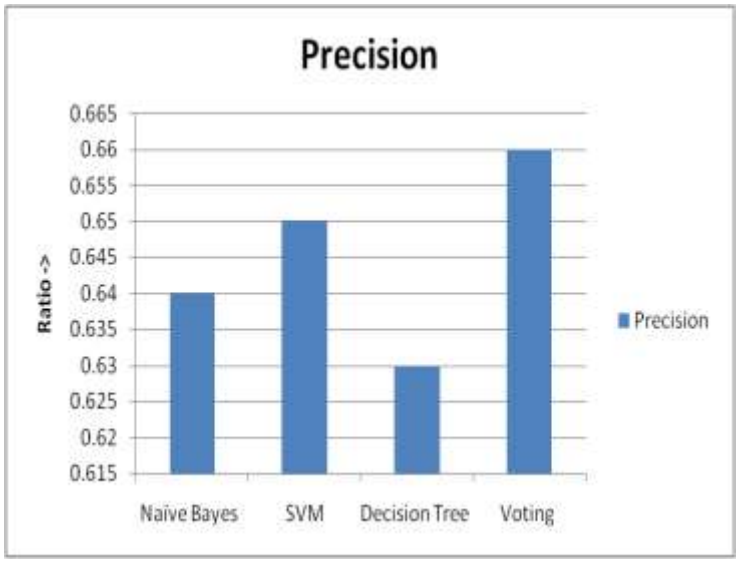

Fig 6 Precision Analysis

As shown in Fig 6, the precision of various classifiers is compared for the email spam detection. The precision classifier has maximum recall which is approx 0.66 as compared to other classifiers

CONCLUSION

The mood of public can be identified by analyzing email spam detection expressed in emails. The voting classifier is developed in this study for email spam detection. The voting classifier combines the three classifiers for analyzing email spam detection. These are naïve bayes, random forest and SVM. These classifiers are combined together for the training of the data. The voting classifier takes input test set and generates predicted result in the form of positive, negative and neutral class. The implementation of recommended and available techniques is done in python. The voting classifier gives maximum accuracy of 72 percent than other classification models of SA.

REFERENCES

[1] Priti Sharma, Uma Bhardwaj, "Machine Learning based Spam E-Mail Detection", 2018, International Journal of Intelligent Engineering and Systems, Vol.11, No.3

[2] M. Deepika, Shilpa Rani, "PERFORMANCE OF MACHINE LEARNING TECHNIQUES FOR EMAIL SPAM FILTERING", 2017, IJRTER

[3] EshaBansal, Pradeep Kumar Bhatia, "A Survey of Various Machine Learning Algorithms on EMAIL Spamming", 2017, International Journal of Advances in Electronics and Computer Science, Vol. 4, No. 3

[4] Dr.SwapnaBorde, Utkarsh M. Agrawal, Viraj S. Bilay, Nilesh M. Dogra, "Supervised Machine Learning techniques for Spam Email Detection", 2017, IJSART, Volume 3 Issue 3

[5] DeepikaMallampati, Nagaratna P. Hegde, "A Machine Learning Based Email Spam Classification Framework Model: Related Challenges and Issues", 2020, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-9 Issue-4

[6] HarjotKaur, Er. Prince Verma, "Survey on E-MAIL Spam Detection Using Supervised Approach with Feature Selection", 2017, International Journal of Engineering Sciences & Research Technology

[7] A.Lakshmanarao, K.ChandraSekhar, Y.Swath, "An Efficient Spam Classification System Using Ensemble Machine Learning Algorithm", 2018, Journal of Applied Science and Computations, Volume 5, Issue 9

[8] W.A. Awad and S.M. ELseuofi, "Machine Learning Methods for Spam E-MAIL Classification", 2011, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1

[9] Madhvi Sharma, Prof. Sumit Sharma, "A Survey of Email Spam Filtering Methods", 2018, Control Theory and Informatics, Vol. 7

[10] PrachiGoyalJuneja, R. K. Pateriya, "A Survey on Email Spam Types and Spam Filtering Techniques", 2014, International Journal of Engineering Research & Technology (IJERT), Vol. 3, No. 3

[11] Samira. Douzi, Feda A. AlShahwan, Mouad. Lemoudden, and Bouabid. El Ouahidi, "Hybrid Email Spam Detection Model Using Artificial Intelligence", 2020, International Journal of Machine Learning and Computing, Vol. 10, No. 2

[12] SushmaL.Wakchaure, ShailajaD.Pawar,GaneshD.Ghuge ,BipinB.Shinde, "Overview of Anti-spam filtering Techniques", 2017, International Research Journal of Engineering and Technology (IRJET), Vol. 4, No. 1

[13] Da Xiao, Meiyi Jiang, "Malicious Mail Filtering and Tracing System Based on KNN and Improved LSTM Algorithm", 2020, IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)

[14] Fahd Aldosari, Sami Bourouis, NizarBouguila, HassenSallay, Khalid M JamilKhayyat, "Infinite Scaled Dirichlet Mixture Models for Spam Filtering via Bayesian and Variational Bayes Learning", 2018,

IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

[15] Bilge KaganDedeturk, BahriyeAkay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm", 2020, Applied Soft Computing

[16] Samira Douzi, Meryem Amar, HichamLaanaya, "Towards A new Spam Filter Based on PV-DM (Paragraph Vector-Distributed Memory Approach)", 2017, Procedia Computer Science

[17] El-Sayed M. El-Alfy, Ali A. AlHasan, "Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm", 2016, Future Generation Computer Systems

[18] VitorBasto-Fernandes, IrynaYevseyeva, Michael T.M. Emmerich, "A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification", 2016, Applied Soft Computing

[19] Maria Habib, HossamFaris, Mohammad A. Hassonah, Ja'farAlqatawna, Alaa F. Sheta, Ala' M. Al-Zoubi, "Automatic Email Spam Detection using Genetic Programming with SMOTE", 2018, Fifth HCT Information Technology Trends (ITT)

[20] KritiAgarwal, Tarun Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization", 2018, Second International Conference on Intelligent Computing and Control Systems (ICICCS)

[21] NishthaJatana, Kapil Sharma, "Bayesian spam classification: Time efficient radix encoded fragmented database approach", 2014, International Conference on Computing for Sustainable Global Development (INDIACom)

[22] Siti-Hajar-Aminah Ali, Seiichi Ozawa, JunjiNakazato, Tao Ban, JumpeiShimamura, "An autonomous online malicious spam email detection system using extended RBF network", 2015, International Joint Conference on Neural Networks (IJCNN)

[23] Vikrant Kumar, Monika, Parveen Kumar, Ambalika Sharma, "Spam Email Detection using ID3 Algorithm and Hidden Markov Model", 2018, Conference on Information and Communication Technology (CICT)

[24] Ahmed I. Taloba, Safaa S. I. Ismail, "An Intelligent Hybrid Technique of Decision Tree and Genetic Algorithm for E-Mail Spam Detection", 2019, Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)