Novel Model for Email Spam Prediction Using Machine Learning

Vinod Kumar

Research Scholar, Department of Computer Science and Application, Kurukshetra University, Kurukshetra, Haryana

Abstract - Email users often receive several hundred spam messages with new content and fresh addresses that are generated automatically by a robot programming tool. Dark-white lists, which are the conventional methods, are useless for preventing spam. To maximise the effectiveness of the spam filtering process, TM (text mining) techniques are applied to emails. The many stages of email spam detection strategies include pre-processing the data, extracting the properties, and classifying the data. The dataset will be cleaned during the pre-processing phase, and features will be extracted to find those that have the greatest influence on the target set. In this stage, the integration of the SVM, NB, and Random Forest classifiers is the main emphasis of the classification process. The proposed system is implemented in Python, and many metrics, including recall, accuracy, and precision, are taken into account while analysing the results. The findings of the suggested model demonstrate significant improvement in email spam prediction.

Keywords - *Email Spam, SVM, Naïve Bayes, Random Forest, Kaggle*

I. INTRODUCTION

A strong, efficient, and discreet form of communication is email. Spammers want to use this type of contact to spread spam. Nowadays, practically all customers have email, which means that they must deal with the spam problem. Spam is a significant problem for both customers and Internet service providers (ISPs). The factors include the acceleration of spam innovation from one perspective and the rate of emergence of electronic communications from another perspective. Email is accessible, which exposes it to a variety of risks brought on by hackers. Email is under grave risk from spam [1], which affects every email client on the planet. Spam refers to unsolicited email and messages that are sent to internet users' inboxes. Thus, sending unrequested data to email boxes can be used to define email spam. The ability to send numerous messages to a large number of customers in a quick and affordable manner is very advantageous for email spammers. It makes this problem applicable to all internet users who often get sporadic email. Spam emails end up being the cause of decreased productivity, take up space in letter boxes, spread bugs, trojans, and materials carrying potentially dangerous data for a particular clientele, thrash the stability of incoming mails [2], and as a result, clients spend their valuable time organizing approaching mail and deleting bothersome messages. Given the large volume of spam mails reaching email inboxes, it may be concluded that spammers collaborate globally and create virtual social networks rather than working in isolation. They assault client emails, large corporations, and even entire states.

Email users often receive several hundred spam messages with new content and fresh addresses that are generated automatically by a robot programming tool. It is nearly impossible to filter spam using traditional methods like dark-white lists (domains, IP addresses, mailing addresses). The effectiveness of email spam filtration may be increased by applying text mining techniques to an email [3]. Additionally, it will be possible to develop topical reliance from topographical features (such as what topics are most prominent in the spam-messages conveyed from the respective nations). Numerous text clustering and classification approaches have been successfully used over the last ten years to address the spam problem. Pre-processing, dimensionality reduction, email categorization, and performance evaluation are just a few of the phases in the pipeline used to predict email spam. Preparing the dataset for data mining is an important step that is done before the mining process. The major goal of this stage is to exclude a few terms from email structures, such as combination words and articles [4], because they are not important for classification. PoS tagging, word stemming, and tokenization are also steps in this process. PoS (Part of Speech) tagging often entails assigning word features to certain PoS (Parts of Speech) based on the review text's determined context. Also carried out in a review text is the tagging of correlation with the surrounding and associated words. POS tags can be used to categorize words as nouns, verbs, adjectives, adverbs, and other types of words. A stemming algorithm converts different word kinds into a single, standardized format. Words from an email's structure are eliminated through the process of tokenization. Additionally, a message is converted into its expressive format. It separates the incoming letter into a number of tokens, or demonstrative symbols. These illustrative symbols are removed from the subject line, header, and body of emails [5]. High dimensionality data is a major problem for both supervised and unsupervised learning. With the present expansion in the size of the existing datasets, this is becoming a significant challenge. Reduced training time, storage space restrictions, and reduced processing overhead are the primary drivers for reducing data size and preserving the number of characteristics at the lowest level feasible. The two main categories of dimensionality reduction approaches are those that rely on feature extraction and those that rely on feature selection. A subset of features is chosen during feature

selection in order to limit the amount of data [6]. A certain cost function is minimized by this approach. In contrast to feature extraction, feature selection does not alter the data and is used to prepare the data before a classifier model is trained. The goal of feature extraction techniques is to condense the amount of the existing feature space into a fresh, smaller feature space. On the basis of linear or nonlinear combinations of features from the real-time dataset, this technique creates new features. Email message classification falls under the category of supervised learning activities. It aims to build a probabilistic model of a function for classifying emails according to emails [7]. A learning algorithm is presented with a set of pre-classified or labelled patterns in the supervised learning of text in email messages, where an entire email dataset serves as one example of a message to be classed. The training set is referred to as this one. Before building a model to be used for testing its effectiveness, certain classified messages from the training set are deleted. The testing set is this collection. Different models are generated using various ways of dividing up the training and testing sets of cases in order to assess the classification accuracy of the finished model [8]. A hyper-plane is used by the SVM algorithm to categorize the data points. Different types of hyper-planes are used to differentiate the data points. This algorithm's main goal is to identify a plan with a maximum range that suggests the greatest distance between the data points in the two groups. To provide some explanation for identifying the data points with greater accuracy, the gap is widened from the margins. Hyper-planes are a type of boundary used to help with decision-making and aid in the differentiation of data points. Various groups are assigned the data points that are available on any side of the hyperplane. Most typically, the number of attributes determines how dimensional the hyper-plane is. If there are three characteristics in the input, the hyper-plane is converted to a two-dimensional (two dimensional) plane [10]. If there are more attributes available, it will be difficult to comprehend the situation. This algorithm is practical and affordable. Due to their suitability for use with the particular data used for training, the Random Forest technique makes use of decision trees (DT). After refreshing the training data, no comparable outcome is present in DT. As they are unable to go back once the data is divided, these factors demand more resources, increase the danger of overfitting, and direct attention to exploring the local optima. The RF (Random Forest) technique addresses these kinds of decision tree flaws. This algorithm combines different models for training the DTs to provide a result. As an example, the permutation strategy for the variable x_i aids in replacing all instances of x_i with random values and identifying the permutation as noise. Thus, the initial relationship between amidx i and the outcome Y is no longer valid. In the interim, the Gini coefficient is used to identify the important factors for reducing a predictor's loss of purity [11]. By using RFCY (random forest cross-validation) to choose the attributes, the data dimensionality is reduced and the variable number is validated. The Bayes theorem, which restricts absolute and conditional probabilities, is the foundation of the class of naive Bayes (NB) classifiers. Probabilities can be connected to the relevant frequency of word appearance in messages in the context of machine learning and spam detection [11]. (i.e., the relative frequency count of words). The following notion is the so-called naive assumption based on the independence of all features with regard to the outcome (i.e., their original class). Despite the fact that this assumption of independence is rarely accurate, naive Bayes classifiers can nonetheless produce a very accurate classification even when there aren't many samples in the training set. Additionally, classifiers from the NB family are regarded as being quick and laid-back. In the last step, the performance of a classifier model is predicted in terms of certain parameters such as specificity, accuracy, sensitivity, and execution time.

II. LITERATURE REVIEW

P. S. Teja, et.al (2021) emphasized on comparing and reviewing the evaluation of certain parameters of supervised ML (machine learning) methods called SVM (Support Vector Machine), RF (Random Forest), DT (Decision Tree), CNN (Convolutional Neural Network), KNN (K-Nearest Neighbor), MLP (Multi-Layer Perceptron), Adaboost (Adaptive Boosting) and NB (Naïve Bayes) algorithm for predicting or classifying the data into spam and authentic emails [12]. The study focused on considering the details or content of the emails, learning a finite dataset available and formulating a mechanism for predicting or classifying the e-mail as spam or ham.

W. Peng, et.al (2018) suggested a new algorithm in order to improve the accuracy of the NB (Naive Bayes) Spam Filter. The real time environment was considered to deploy the suggested algorithm. A novel addition coded to the current servers to enhance the accuracy of Spam Server Spamassassin. This approach focused on analyzing the specific aspect of spam emails due to which various challenges were occurred for individuals and companies. The results revealed that the suggested algorithm assisted in mitigating the amount of spam emails whose misclassification was done as authentic email. Moreover, the suggested algorithm was capable of maximizing the accuracy of email sorting as well as proved as a valuable addition to the current systems.

H. V. Bathala, et.al (2021) introduced a technique in to consider the text of the body of the email as well as handle the embedded phishing URLs and spam images which were contained in email [14]. Diverse recent ML (Machine Learning) techniques adopted for classifying the emails and a structured procedure was put forward for recognizing the spams. The effective ML algorithm was selected using a lazyPredict library. The classic datasets were employed to analyze this technique and indicated that the adopted techniques were useful to recognize the spams and prevent the zero-day assaults. The results depicted that the Stacking

offered an accuracy of 97% to detect the phishing URLs. Moreover, the MLP (Multilayer Perceptron) attained accuracy of 97% to detect the email spams.

S. M. Y. Abouelseoud, et.al (2022) developed a DL (deep learning) algorithm to illustrate the enhancing efficacy over the traditional methods [15]. The experiments were conducted on 3 datasets in which the content-based attributes comprised. The developed algorithm helped in differentiating 3 classes to filter the general spam. This algorithm was validated and tested with regard to diverse parameters. The practicality of the developed algorithm was proved in the diverse applications and it offered superior accuracy and efficiency. The results exhibited that the developed algorithm performed more accurately as compared to the other methods.

S. Priya et.al (2020) investigated a STEPD-KELM framework for detecting the CD (concept drift) on the basis of computing the variation in the email content distribution [16]. STEPD (Statistical Test of Equal Proportion) method was constructed for determining the criteria of CD for all unknown emails which aided the filtering method to recognize the occurrence of the spam. Furthermore, the instances were classified into two classes: spam and ham using KELM (kernel extreme learning machine). Enron dataset applied to test the investigated framework in the experimentation. The outcomes of experiments revealed that the investigated framework yielded the precision up to 93.78%, recall around 96.54%, and accuracy up to 95.33%.

Y. Zhang, et.al (2019) formulated a GTRS (Game-Theoretic Rough Set) approach in order to filter the email spam so that an appropriate three-way technique to filter the email spam was generated for illustrating a tradeoff amid the accuracy and coverage [17]. This work concentrated on analyzing the GF (Game formulation) and RL (repetition learning) of formulated approach. UCI spam dataset executed in the experimentation. A comparative analysis was performed on the formulated approach against the traditional technique. The experimental outcomes reported the adaptability of the formulated approach for enhancing the coverage level in considerable manner.

S. K. Sonbhadra, et.al (2020) recommended a new dynamic spam filter called multi-tier method in which the changes were taken in account in the interests of users with time when the spam activities were handled [18]. The IS (intention-based segmentation) model was executed for comparing dissimilar segments of text documents. First of all, the major intend of the recommended method was to split the email content into segments. For this, POS (part of speech) tagging was implemented on the basis of voices and tenses. After that, the hierarchical clustering adopted to cluster the segments and compare them via vector space model. In addition, the change in the interest of the user was recognized after detecting the concept drift. In the end, this method classified the emails into diverse classes. Enron dataset applied to simulate the recommended method. The results exhibited that the recommended method had generated optimal outcomes.

S. J. S. Daisy, et.al (2021) established a Hybrid method in which the NB (Naive Bayes) algorithm was put together with the MRF (Markov Random Field) [19]. The occurrence and outline of values in a dataset were verified using NB algorithm and a probabilistic classification operation was executed. The Bayes Theorem was deployed to classify the emails as spam or ham. HMF algorithm assisted in simulating the statistical action of spam. The function vectors were built by dividing the email into attributes. The weighting of these attributes was done for compensating the inter-word dependence in the learning algorithms. The results demonstrated that the established method was effective in detecting the spam concerning accuracy and time utilization.

N. Mageshkumar, et.al (2022) suggested a new technique for enhancing the accuracy of NB (Naive Bayes) filter for detecting the text alterations and correctly classifying the emails as spam or ham [20]. The semantic based, keyword based, and ML (machine learning) algorithms were integrated for maximizing the accuracy of the suggested algorithm. In addition, this algorithm aimed at finding the association of length of the email with the spam score, exhibited that the Bayesian Poisoning was actually a real phenomenon and the spammers utilized it mostly.

III. RESEARCH METHODOLOGY

A voting classification method is planned on the basis of various classifiers. There are two categories of voting methods namely soft and hard voting. The initial category is employed to assign weights to every classification model for voting. The model is trained using several classifiers. The trained algorithms offered input to the voting method in which the input test data was deployed and the results are extracted. This kind of method is called Supervised ML (machine learning) algorithm. Such a technique assists in accomplishing the regression, classification and outlier discovery. The concept of a hyperplane has great influence on these techniques while classifying the data. A hyperplane deploys to represent the subspace of a vector space. Its vector space consumes less than one dimension. The most optimal hyperplane helps in detecting the larger difference amid two separate planes. The categories of NB (Naive Bayes) algorithms are depending on the Bayes' theorem to restrict the complete and conditional probabilities. The ML and techniques of detecting spam are utilized to associate the probabilities with the related frequencies of word presence in messages. The subsequent notion is the hypothesis on the basis of independence of all attributes for which the output is considered. This supposition is precise and this algorithm is able to classify the data correctly even in the case of absence of various instances in the training data. Moreover, the algorithm of these category performed quickly and reliably. Moreover, SVM (support vector machine) is another

algorithm which aids in detecting the outlier. This algorithm aims to discover the largest distance among data points related to distinct classes to investigate the separation hyperplane in the presence of labeled dataset. Hard-margin and soft-margin are its two categories. Different from K-NN (K-Nearest Neighbor), this algorithm is applicable in areas having high dimensionality. The amount of features is increased to partition the data points. The support vectors are defined as the points which are present nearer to the hyperplane. A decision boundary is another term assigned to this hyperplane. It is effective to split the components of distinct groups.

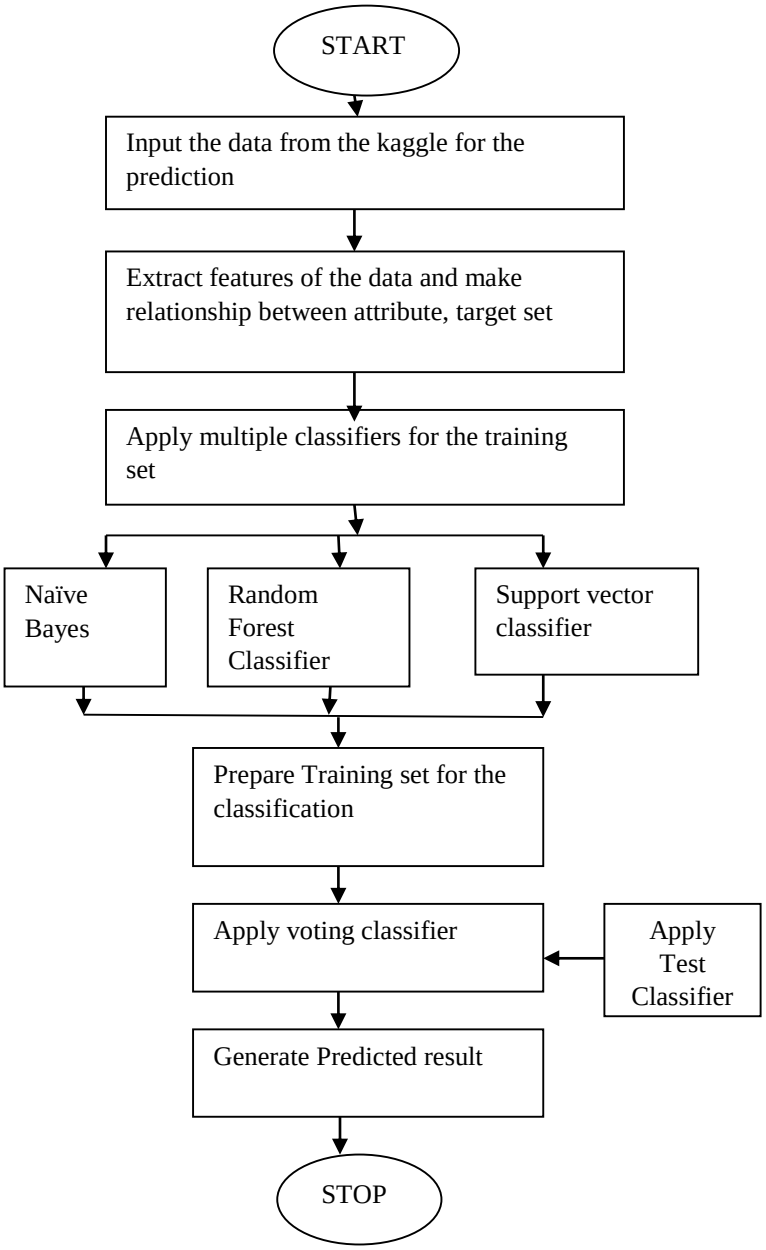

Figure 1: Proposed Flowchart

IV. RESULT AND DISCUSSION

This study is connected to the forecast of email spam. The process of foretelling spam in email involves a number of steps, including pre-processing the data, extracting the properties, and classifying the data. The material will be cleaned in the pre-processing stage, and the data will be categorised into several classes in the classification stage. To assess the model's dependability, the new model is compared to more established models like SVM, KNN, and Random Forest.

Table 1: Dataset Details

Number of Attributes	2
Number of Instances	5174
Number of Training Samples	2914
Number of Test Samples	727
Data Division set	60 Percent

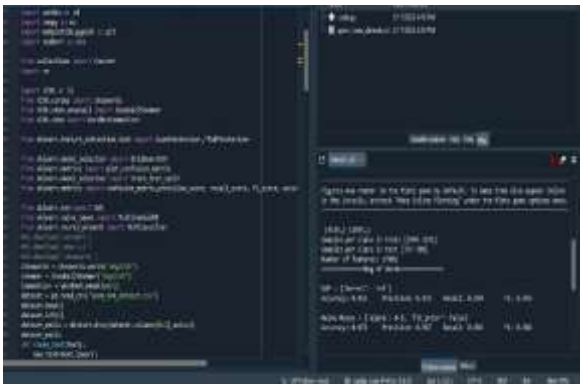

Figure 1: Deployment of the Model Built

Figure 1 shows how the suggested model, which is a mix of several classifiers, is put into practice. The suggested Model increases the email spam detection's precision.

Table 1: Comparative Study

Model Name	Accuracy	Precision	Recall
SVM	76.89 percent	78.90 percent	76 percent
KNN	74.56 percent	74 percent	74.2 percent
Random Forest	80.12 percent	80 percent	80 percent
Proposed Model	90.12 percent	91 percent	91 percent

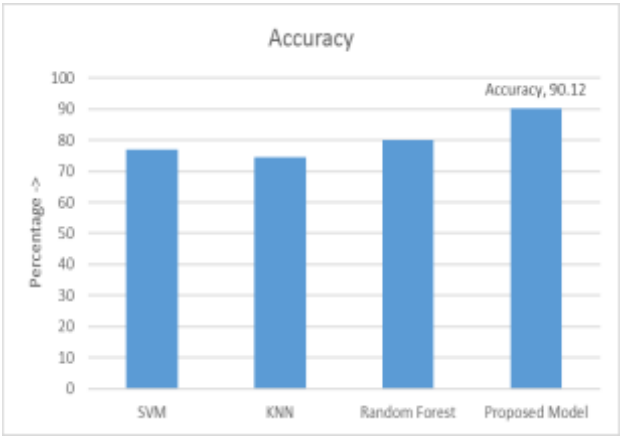

Figure 2: Analysis in the context of accuracy

Figure 2 compares the performance of the proposed hybrid model for email spam prediction with that of the Support Vector Machine, KNN, and Random Forest. The investigation showed that the proposed method could predict email spam with an approximate accuracy of up to 90%.

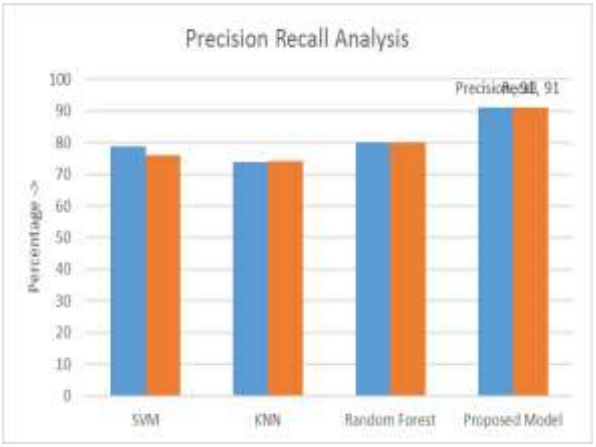

Figure 3: Analysis based on Precision Recall

In terms of precision-recall values, Figure 3 compares the suggested strategy to the Support Vector Machine, KNN, and random forest. In order to identify email spam, this method blends NB, SVM, and RF.

V. CONCLUSION

According to this article, predicting email spam is a challenging machine learning task. Several techniques have been developed in previous years to identify email spam based on tokenization, word polarity, etc. This study presented an ML method based on different phases for email spam identification. To pre-process the data, extract the features, and categorise the data, the phases are carried out. The classification method that is presented in this paper combines Random Forest, NB, and SVM. The findings showed that the new approach can attain accuracy levels of up to 90%. The precision and recall percentages will be 91% and 91%, respectively.

VI. REFERENCES

[1] Dr. Swapna Borde, Utkarsh M. Agrawal, Viraj S. Bilay, Nilesh M. Dogra, “Supervised Machine Learning techniques for Spam Email Detection”, 2017, IJSART, Volume 3 Issue 3

[2] Deepika Mallampati, 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

[3] A. Lakshmanarao, K. Chandra Sekhar, Y. Swath, “An Efficient Spam Classification System Using Ensemble Machine Learning Algorithm”, 2018, Journal of Applied Science and Computations, Volume 5, Issue 9

[4] Apurva Taunk, Srishty Bharti, Sipra Sahoo, “An Ensemble Method for Spam Classification”, 2020, International Journal of Scientific & Technology Research Volume 9, Issue 02

[5] Megha Rathi, Vikas Pareek, “Spam Mail Detection through Data Mining – A Comparative Performance Analysis”, 2013, International Journal of Modern Education and Computer Science, Volume 12, PP. 31-39

[6] Hanif Bhuiyan, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas & Jinat Ara, “A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques”, 2018, Global Journal of Computer Science and Technology, Volume 18, Issue 2

[7] Harjot Kaur, Er. Prince Verma, “Survey on E-mail Spam Detection using Supervised approach with Feature selection”, 2017, International Journal of Engineering sciences & Research technology

[8] G. Vijayasekaran, S.Ros, “Spam and Email Detection in Big data Platform using Naives Bayesian classifier”, 2018, International Journal of Computer Science and Mobile Computing, Vol.7 Issue. 4, pg. 53-58

[9] Yukti Kesharwani, Shrikant Lade, “Spam Mail Filtering Through Data Mining Approach –A Comparative Performance Analysis”, 2013, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 9

[10] Prabha Pandey, Chetan Agrawal, Tehreem Nishat Ansar, "A Hybrid Algorithm for Malicious Spam Detection in Email through Machine Learning", 2018, International Journal of Applied Engineering Research, Volume 13, Issue 24 pp. 16971-16979

[11] Deepika Mallampati, 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

[12] P. S. Teja, C. Amith, K. Deepika and D. K. S. Raju, "Prediction of Spam Email using Machine Learning Classification Algorithm", International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 9, no. 4, pp. 831-837, 2021

[13] W. Peng, L. Huang, J. Jia and E. Ingram, "Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection," 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 2018, pp. 849-854

[14] H. V. Bathala, P. V. N. P. Srihitha, S. G. R. Dodla and A. Pasala, "Zero-Day attack prevention Email Filter using Advanced Machine Learning," 2021 5th Conference on Information and Communication Technology (CICT), 2021, pp. 1-6

[15] S. M. Y. Abouelseoud and M. Mikhail, "Efficient spam and phishing emails filtering based on deep learning", Computer Networks, vol. 18, no. 6, pp. 1231-1245, 15 February 2022

[16] S. Priya and R. Annie Uthra, "An Effective Concept Drift Detection Technique with Kernel Extreme Learning Machine for Email Spam Filtering," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 774-779

[17] Y. Zhang, P. Liu and J. Yao, "Three-way Email Spam Filtering with Game-theoretic Rough Sets," 2019 International Conference on Computing, Networking and Communications (ICNC), 2019, pp. 552-556

[18] S. K. Sonbhadra, S. Agarwal, M. Syafrullah and K. Adiyarta, "Email classification via intention-based segmentation," 2020 7th International Conference on Electrical Engineering, Computer Sciences and Informatics (EECSI), 2020, pp. 38-44

[19] S. J. S. Daisy and A. R. Begum, "Smart material to build mail spam filtering technique using Naive Bayes and MRF methodologies", Materials Today: Proceedings, vol. 17, no. 12, pp. 8442-8452, 27 May 2021

[20] N. Mageshkumar, A. Vijayaraj and A. Sangeetha, "Efficient spam filtering through intelligent text modification detection using machine learning", Materials Today: Proceedings, vol. 5, no. 2, pp. 213-222, 30 May 2022