IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A DETAILED EVALUATION OF JUNK MAIL CLASSIFIERS METHODS

[1]Gaurav Dahiya, [2]Shivkant, [3]Kirti Bhatia

[1]Student, [2]Assistant Professor, [3]Assistant Professor

[1]Computer Science & Engineering,

[1] Sat Kabir Institute of Technology and Management, Haryana, India

Abstract: In this paper, we examined the issue of junk mail, which is a significant Internet-related concern. There is now a need for a robust anti-spam filter due to the increase in uncalled mass email, also known as junk Mail. Nowadays, machine learning (ML) procedures are used to filter emails that are spam effectively and opportunistically. In this paper, we've examined some of the popular machine learning (ML) techniques, including rough sets, Bayesian classification, SVMs, k-NN, ANNs, and artificial immune systems, and their applicability to the problem of spam email taxonomy. On the basis of the amount of E-mail Assassins, we have supplied descriptions of the steps and the variations in how they are carried out.

Index Terms **– Junk Mail, Machine learning Procedures, Artificial Intelligence.**

I. INTRODUCTION

We now confront new hazards as a result of the growing popularity of the Internet, such as virus attacks and spam, or unsolicited bulk emails that are sent for commercial gain. These e-mails use up storage space, communication bandwidth, and time. Despite numerous firewalls, this problem has been becoming worse and worse. Fewer than 50% of emails are spam, costing internet users a fortune each year, according to a previous survey. In this case, the best method for thwarting spam at any time may be an automatic e-mail filtering system. Consequently, spammers and anti-spammers are engaged in a fierce battle. In the past, spam was filtered by identifying and blocking spam email addresses, or by filtering messages containing particular keywords or advertising. The use of multiple sender IP addresses or appending/suffixing random symbols to the beginning/end of the message's subject line are just two of the many strategies spammers have started to utilise to trick screening mechanisms (Cormack, 2011).

Knowledge engineering (KE) and machine learning (ML) are the most often used techniques for spam mail filtering. In KE-based approaches, we specify a set of criteria to distinguish between legitimate emails and spam emails. The user or the software provider who provides the rule-centered spam-filtering software can provide this set of rules. It is always impossible or potentially inconvenient for many people to regularly renew and maintain these norms, which is necessary for success. It has been determined that machine learning (ML) approaches are superior to knowledge engineering approaches since they do not require the specification of any directives (Guzella, 2011). Instead, a collection of pre-specified email posts serve as the training samples for machine learning techniques. Following that, a specific algorithm was used to extract the categorization rules from these email posts. Recently, a number of machine learning techniques that can be utilised for email filtering have received extensive reviews. These techniques include the artificial immune system, J48 classifier, support vector machines (SVM), Naive Bayes, and neural networks. This effort placed a strong emphasis on the analysis of spam mail filtering and preserving regular emails (Jayant, 2021).

II. RELATED WORK

Email must be filtered in order to be divided into ham and spam. Through the identification of distinctive qualities, the authors (Mohamad, 2015) have suggested a spam email filtering approach that divides emails into legitimate and unwanted ones. They utilised TF-IDF and rough set theory after pre-processing the dataset's (English and Malay email) descriptions. They then used a machine learning methodology to categorise data and achieved excellent results. (Harisinghaney, 2014) have suggested a brand-new ML-based technique for classifying email data. The algorithmic implementation includes KNN and Naive Bayes algorithms and presents practical results in cases where algorithms are used to pre-processed datasets.
Also, the researchers (Aljarah, 2015) have designed an email filtering strategy that is ontology-based. The J48 decision tree-centered method was utilised to categorise the used dataset. A RDF linguistic-centered ontology was created by Jena in order to test the results obtained after categorization.

III. COMPARATIVE ANALYSIS OF STANDARD WORK

Table 3.1 A detailed analysis of different machine learning methods for junk mail detection.

Author	Research Outcomes	Research DIFFERENCE
Guzella T, Caminhas W. M., 2000	The writers of this work have provided a comprehensive overview of recent advancements in the application of ML techniques for spam filtering. They have placed a strong emphasis on both text- and image-centered approaches. Instead than seeing spam filtering as a standard categorization issue, they have emphasised the importance of reflecting specific aspects of the issue, like perception significance for the conception of various filters.	They have examined the difficulty in updating a classifier that gives the bag-of-words example and a crucial change in older naive Bayes models. They have summarised the substantial advancement in this area and identified additional qualities that still need to be uncovered, especially in more plausible estimating scenarios. .
Levent Özgür, Tunga Güngör, and Fikret Gürgen, 2004	Based on ANN and Bayesian classifiers, the authors have suggested aggressive anti-spam cleaning methods for various vernaculars in general and specifically for Turkish. Both the processes and the related components are adaptable. The main module focuses on morphology, and the following module classifies the emails based on their roots. Single layer and multi layer perceptron (MLP) ANN structures have been taken into consideration, and the inputs to the networks are managed by binary and probabilistic prototypes. For the categorization of models, they have employed three different strategies: binary, probabilistic, and advance probabilistic.	The MM and LM components make up the anti-spam filtering system as it has been defined in this work. A Turkish morphological investigation approach has been developed. Turkish is an agglutinative language since a given word in it may be related to an idiom composed of many terms. As a result, morphological analysis of Turkish is more challenging than morphological analysis of languages like Hindi. They have employed the concept of shared information in order to identify the origin terms that may be used as the components of the categorization technique. A set of intense words that can be used for categorization have been referred to as the aspect vector. The potential terms were initially detected.

Enrico Blanzieri , Anton Bryl, 2007	The authors of this study examined a spam filter example that was based on the SVM-NN classifier theory. The ideas of SVM and kNN have been integrated. A SVM prototype that has been trained for these k samples is then used to mark the unknown sample after the classifier has first found the k closest marked messages to the message to be tagged. The results of the authors' comparison of SVM-NN with SVM and k-NN are shown.	They have employed Euclidean metric in their trials to determine the linear kernel for SVM and the neighbouring neighbours. An unordered list of strings with gaps separating each string served as the basis for all messages. A binary feature of a communication was reflected when a specific token appeared in a specific part of a message, particularly in the header. After that, the training dataset's most recurrent characters were picked and used. Consequently, a vector of binary descriptors has been used to characterise each message. The researchers evaluated a spam filter that is learning-focused. The suggested approach significantly outperforms SVM in the feature domain's minuscule dimensions.
Mawuena Glymin and Wojciech Ziarko, 2007	In this paper, approaches for spam recognition using probability evaluation tables and analytical data modelling are fundamentally summarised. The main focus is on the solution's presentation, that combines straightforward techniques with specific heuristics to use the VPRSM rough set methodology to generate complete approximate estimates of spam and legitimate e-mails. The use of VPRSM for developing a smart worker for spam filtering was the subject of experiments. They have looked into the viability of using the VPSRM probability tables hierarchy as a spam recognition filter.	A training step and a categorization stage have been assigned to the operational spam recognition system. The pool of pre-categorized emails uses the rough set centred machine learning, which was employed in the training phase. A hierarchy of trained decision tables was used during the categorization stage to predict the preferred class of email recipients. They examined the determination prototype using emails from the Hotmail platform that were acquired for training purposes.
Hsu Wei-Chih, Tsan-Ying Yu, 2009	The SVM is a well-known data mining technique for categorization and has been successfully applied in a number of real-world situations. The SVM training procedure specifically involves parameter selection that has an impact on classification results. However, knowledge or straightforward grid searches are typically used to identify the selection parameter in SVM. They advocated the Taguchi technique, which was used to develop the SVM-based	The Taguchi technique was used by the authors to identify the ideal combination of the two SVM coefficients (C,). The coefficients of SVM were taken into consideration as influencing elements when developing a spam filtering prototype. After choosing the SVM coefficients, they compared grid search to the classification results and validated them. The SVM is a powerful supervised learning model built on the organised risk-lowering standard

	Spam Filtering prototype and grid search. The realisation of orthogonal arrays without repetition is straightforward. To demonstrate the effectiveness and potential of the technique, a physical world mail dataset was selected. The results of the experiment show that the Taguchi technique can find the working prototype with high classification accuracy and strong robustness.	from mathematical learning principle. SVM's implementation of the text categorization process is astounding.
Ibrahim Eldesoky, 2009	The immune system's natural response has been used to trigger an artificial immune system (AIS) prototype. The AIS prototype that the authors have suggested satisfies the spam filtering process. The email messages' information have been used for both testing and training operations. The email's words are weighed and used to estimate the level of empathy between an antigen and an antibody. They have started a learning phase to reward the cells that correctly recognise the spam email. Due to the condensed number of sensors, negative assortment was used during the training period rather than clonal assortment, which resulted in better testing execution. The fraction of training group components is significantly lower than other techniques when compared, which supports the enhanced enactment. The method was tested against the vast amounts of spam and regular mail received worldwide.	The main component of the spam immune system is the detector or sensor, and they are referred to as antibodies (AB). The subjects of email posts must be verified by the ABs. They first calculated the native weight of each word in the email in order to convert it into an AB. By adding up each word's occurrences relative to the amount of words in the email, the native weights were calculated in a consistent manner. Instead of totals of the similar words in the two emails, native weights are used to support the predicted affinity assessment approach. They then compared this affinity to a predetermined value to determine the type of email by comparing it to that value. Even though the artificial prototype has been trained to recognise both unwanted and spam elements, it discards the cells that incorrectly recognise new antigens as the testing process progresses.
Loredana F., Camelia L., Rodica P., 2010	In order to categorise the emails, this work comprises a novel spam identification filter that combines a number of features, many of which are heavily reliant on word frequency. The suggested design has been divided into two major modules: one for gathering information and retrieval from all communications, and another for	The kNN method, which is complemented by a group of features mined from the attributes and contents of the email, was used by the authors to classify the messages. The trained set has been reproduced to the best possible size, and several experiments have determined the best method for class dissemination.

	categorising emails and analysis of the findings. In the beginning, an email has been chosen to gather the data for analysis. To draw spammers, they have created spam baits. They have analysed the message based on many variables including message length, the number of answers, and word frequency in order to detect the characteristics. The representative set of the data collection is represented by an email and these extracted attributes. The researchers combined a curbing approach with the tokenization technique for the message's contents.	The suggested mechanism accomplishes the continual renewal of the dataset and the record of the most frequently occurring words found in emails. Additionally, they offered a feedback option for unclassified messages and advised that it be done at a set rate subject to the availability of the necessary resources.
Tiago A. Almeida · Jurandy Almeida · Akebo Yamakami, 2011	Here, the researchers have taken into account the implementation of various term-selection techniques employing various independent Naive Bayes (NB) spam filter prototypes. They carefully planned the experiments to confirm the conclusions that were obtained through statistical analysis. Additionally, they have conducted research on the metrics typically used to judge the effectiveness of spam filters. The advantages of using the MCC as a standard of enactment have also been looked into.	First, the authors demonstrated a Bayesian decision scheme-based classifier's implementation assessment of several term-based techniques in the field of spam filtering as dimensionality decreased. They have completed the analysis of the results obtained by several Naive Bayes spam filters that were used to classify emails from six well-known, actual, public, and sizable email datasets. Additionally, they suggested the MCC as the measuring factor for estimation, which offers a more consistent assessment of the forecast than TCR, especially if the two groups have different sizes.
Yuanchun Zhu and Ying Tan., 2011	The researchers have suggested a local concentration (LC) inspired feature mining approach for spam filtering, which was stimulated by the biological immune system. By transforming each message region into an analogous LC component, this approach can extract location-related information from communications. They have implemented two LC technique methods using sliding windows with fixed and variable lengths. A generic LC model has been designed to include the LC method into the complete	They have carried out multiple experiments using 5 standards masses using the cross confirmation technique to evaluate the projected LC system. Three term selection methods have been shown to complement the LC methodology well. The LC methodology has better implementation in terms of both precision and measurement when compared to the existing bag-of-words technique and the universal concentration centre method. They have also confirmed that the LC approach is effective for messages of various lengths.

	practice of spam filteration. Two kinds of detector groups have been producedthrough term selection approaches and a clear penchant threshold. Afterward a sliding window has been accepted to allocate the message into separate regions. Once segmentation of the Message is complete, next they have calculated the strength of detectors and measure the aspect for every native region. Conclusively, they have combined all the aspects of local areas as a aspect vector of the message.	
Noemí Pérez-Díaz, David Ruano-Ordás, José R. Méndez, Juan F. G., Florentino F., 2012	The authors have reviewed and combined the preceding methods and new substitutes for smearing the rough set (RS) mechanism over the spam filtering space through describing the three distinct rule implementation techniques such as MFD), LNO and LTS. Keeping in mind the aim of properly evaluating the correctness of the anticipated procedures, they precisely solve and review important queries for suitable prototype approval such as corpus assortment, pre-handling and representative problems, and distinctive exact standard processes. the testing done using various implementation strategies to select the best decision rules created by RS. Their predicted methods outperformed other well-known spam filtering techniques including SVM, Adaboost, and several Bayes classifier types.	This article offers a thorough review of the applications of RS as a primary classifier for commercial spam filtering. They provided and looked into many methods using RSs, along with a realistic analysis of their uses, benefits, and downsides. Following examination, they came to the conclusion that the majority of the earlier studies used corpora and insufficient pre-handling to provide a concept. In light of this, the authors conducted a new research for a sizable, fresh, and unmanaged corpus that was provided by the SpamAssassin team. The RS-centered approaches are consistently a suitable replacement for the Naive Bayes classifier, SVM, and Adaboost algorithm, as confirmed by the results obtained. When combined with LNO and LTS, the suggested MFD heuristic achieves the highest level of precision.
Yazdan Jamshidi, 2016	This paper demonstrates the applications of Interval's Number KNN (INKNN) for spam filtering. This technique was later described as a lattice data set enlargement of the KNN technique. An IN was being used to display a population of spam emails. The developed classifier was then used to distinguish spam emails from legitimate ones.	This study presents a NN classification method for spam filtering built on the concepts of lattice and probability-explained interval numbers. Modelling both ambiguous data and other types of lattice-ordered data is a real-world advantage of the lattice approach. The suggested approach works with a variety of data types. Both locations and intermissions can be

	They have conducted extensive experiments over the public SpamAssassin corpus to gauge the effectiveness of INKNN. The results show that, in comparison to other earlier advanced ML approaches, the INKNN is capable of accomplishing the advanced enactment.	managed by it. Since the suggested method involves a quick learning process, it can be applied to a variety of situations where the amount of data is so great that thorough examinations take a long time. Utilising an interval number has the main advantage of allowing for the lodging of imperfect data. The results of the experiment confirm the effectiveness of the working model they had planned.
Ali Shafigh Aski, Navid Khalilzadeh Sourati, 2016	Through a multilayer perceptron prototype, this work defines three machine learning (ML) techniques to discriminate between spam and hams with low mistake rates and high competence. Decision tree (DT) classifier, multilayer perceptron (MP), and NB classifier are a few widely used techniques. These techniques are helpful for training datasets that are either hams or spam. Finally, they have reviewed and examined the measured method findings in relation to the intended prototype.	The suggested approach is founded on standards for acceptable recording in the context of instruction competency. Three different types of instructions were given: (1) information analysis of the email header, (2) keyword counting, and (3) important message content.
Priti S., Uma B., 2018	The goal of this work is to propose a machine learning (ML) based fused bagging approach for spam email recognition by combining the two ML methods (NB and J48 (DT)). Here, the dataset has been divided up into various groups and given to all algorithms as input. They carried out several trials, comparing the findings in terms of precision, memory, exactness, f-measurement, and true and false negative/positive rates. Naive Bayes was used in one experiment, and J48 techniques were used in the other. The hybrid bagged approach was used to conduct one more experiment. The hybrid bagged technique of spam mail detection system achieved a total precision of 87.5%..	A multcast decision tree classifier, the J48 approach is based on the concept of entropy. Using the training dataset, it creates decision trees. The DT is produced by J48, and the categorization of the new data value depends on the attribute values of the training data. It adheres to the concept of grouping the data into many categories so that every aspect of its quality can be used to inform decisions. The procedure repeats until each data attribute has been controlled and categorised.

Abdul J. S., Asif K., Bharanidharan S., Sami A. , Krishnan K., Mirjam J. and Friso D., 2019	The use of spam emails for non-personal, unwanted moneymaking, or malicious purposes is also common. These emails are sent and forwarded with the intention of upsetting either a person, a business, or a group of people. In addition to being promotional, the emails may be linked to websites that hold phishing or malware that can steal personal information. The authors of this paper investigated the effectiveness of applying an NSA approach for spam filtering and inconsistency recognition. The proposed technique has a low rate of false recognition and a high enactment.	The suggested model is trained by creating a memory of spam emails' prior behaviour. Because the model has been educated against a particular activity, it prohibits the same sort of behaviour for incoming messages in the future. The process in question is called negative selection (NS). The self- and nonself discerning behaviour of the mammalian learning immune system served as the foundation for the NSA's plan. The goal is to create patterns that do not correspond to or are equal to an existing body of readily accessible patterns in order to create a prototype of irregularities, variations, or oblivious facts. The NSA analyses data on its own and other people's behaviour to identify gaps between regularity and irregularity.

Fig. 1 shows a working model of Junk Mail Detection using Machine Learning Method.

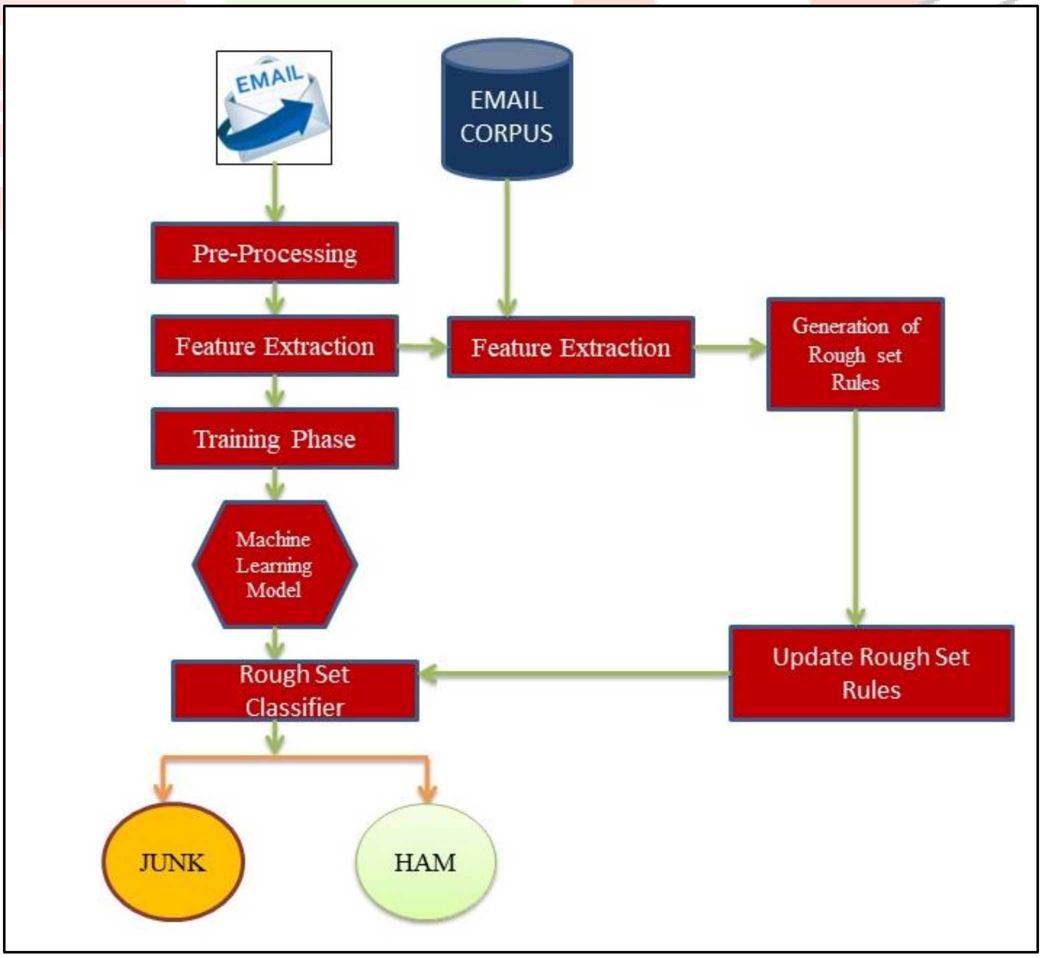

Fig. 1: Rough Set Based JUNK Mail Identification

IV. CONCLUSION

In this work, we have reviewed certain well konwn machine learning techniques and their suitability to the issue of spam e-mail categorization. We have presented the research gaps between the machine learning approaches towards spam e-mail filtration. We have presented brief descriptions of the popular machine learning methods like Naïve Bayes classifier, KNN, ANN, rough sets, and artificial immune systems. In context of precision, we conclude that the Naïve bayes and rough sets approaches are very satisfying in enactment amid the other approaches. There is a need of extra research to increase the enactment of the Naïve bayes and Artificial immune system either by amalgam system or by resolution of the feature dependence problem.

The use of machine learning classifiers by several academicians to address the spam problem was highlighted. Studies show that spam communications have changed over time to get past sieves. It was looked into how spam emails are screened as well as the fundamental design of an email spam sieve. The study examined some publicly available datasets and implementation techniques that could be used to evaluate spam sieve effectiveness. The challenges faced by ML procedures in successfully addressing the spam problem were addressed, and comparative evaluations of ML strategies found in the literature were carried out. With spam sieves, we have come across a number of unresolved research problems. Overall, the quantity and variety of the materials we looked at show that this industry has had and will continue to experience great expansion. Following the explanation of the unresolved problems in spam screening, additional study is needed to improve the effectiveness of spam sieves. The development of spam sieves will therefore continue to be a focus for academics and business consultants researching ML techniques for practical spam screening. We anticipate that our study will act as a springboard for excellent research in spam filtering by researchers using ML, deep learning, and deep adversarial learning processes.

.

REFERENCES

Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.

[1] Cormack, L., Gordon, S., Mark. C. 2011. Efficient and effective spam filtering andre-ranking for large web datasets Information Retrieval, Springer Netherlands, 1-24.

[2] Guzella, T. S. and Caminhas, W. M. (2009) "A review of machine learning approaches to Spam filtering." Expert Syst. Appl..

[3] Mohamad, M. and. Selamat A. (2015). An evaluation on the efficiency of hybrid feature selection in spam email classification. In: Proc. of 2015 International Conference on Computer, Communications, and Control Technology (I4CT), Kuching, Sarawak, Malaysia, 227-231.

[4] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora (2014). Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm, In: Proc. of 2014 International Conference on Optimization, Reliabilty, and Information Technology (ICROIT), Faridabad, Haryana, 153-155, India,

[5] S. Youn, and D. McLeod (2007) Efficient spam email filtering using adaptive ontology. In: Proc. of Fourth International Conference on Information Technology, Las Vegas, NV, USA, pp.249-254,

[6] H. Faris, and I. Aljarah, (2015). Optimizing feedforward neural networks using Krill Herd algorithm for e-mail spam detection, In: Proc. of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, 1-5,

[7] Guzella T, Caminhas W, (2000). A review of machine learning approaches to spam filtering. Exp Syst Appl, vol. 36, issue 7, 10206–10222.

[8] Levent Özgür, Tunga Güngör, and Fikret Gürgen, (2004). Spam Mail Detection Using Artificial Neural Network and Bayesian Filter", in International Conference on Intelligent Data Engineering and Automated Learning, 505-510.

[9] Enrico Blanzieri and Anton Bryl, (2007). Instance-Based Spam Filtering Using SVM Nearest Neighbor Classifier, American Association for Artificial Intelligence, www.aaai.org, 441-442.

[10] Mawuena Glymin and Wojciech Ziarko, (2007). Rough Set Approach to Spam Filter Learning", in RSEISP '07: Proceedings of the international conference on Rough Sets and Intelligent Systems Paradigms, 350–359.

[11] Hsu Wei-Chih and Tsan-Ying Yu, (2009). E-mail Spam Filtering Using Support Vector Machines with Selection of Kernel Function Parameters, in Fourth International Conference on Innovative Computing, Information and Control, 765-767.

[12] M. H. Haggag and I. E. Fattoh, (2009). Artificial Immune System for Spam Filtering, IJICIS, 9(2), 117-129.

[13] Loredana Firte, Camelia Lemnaru and Rodica Potolea, (2010) "Spam Detection Filter using KNN Algorithm and Resampling", in International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, . 27-33.

[14] Tiago A. Almeida, Jurandy Almeida and Akebo Yamakami, (2011.) Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers, J Internet Serv Appl, 183-200.

[15] Yuanchun Zhu and Ying Tan, (2011). A Local-Concentration-Based Feature Extraction Approach for Spam Filtering, IEEE Transactions on Information Forensics and Security, 6 (2), 1-25.

[16] Noemí P., David R., José R. M., Juan F. G. and Florentino F., (2012). Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification, Applied Soft Computing, 3671–3682.

[17] Yazdan Jamshidi, (2016). A nearest neighbour classifier based on probabilistically/possibilistically intervals' number for spam filtering", Int. J. Soft Computing and Networking, (1), 4-16.

[18] Ali Shafigh Aski and Navid Khalilzadeh Sourati, (2016). Proposed efficient algorithm to filter spam using machine learning techniques, Pacific Science Review A: Natural Science and Engineering, 145-149.

[19] Priti Sharma and Uma Bhardwaj, (2018). Machine Learning based Spam E-Mail Detection, International Journal of Intelligent Engineering & Systems, 11(3), 1-10.

[20] Jyh-Jian Sheu,Yin-Kai Chen, Ko-Tsung Chu, Jih-Hsin Tang and Wei-Pang Yang, (2016). An intelligent three-phase spam filtering method based on decision tree data mining, Security and communication networks, 9 (17), 40130-4026.

[21] Shradhanjali and Toran Verma, E-Mail Spam Detection and ClassificationUsing SVM and Feature Extraction, nternational Journal of Advance Research, Ideas and Innovations in Technology, 3,(3), 1491-1495.

[22] Jayant Batra, Kirti Bhatia, Rohini Sharma, Shalini Bhadola. (2021). An Overview on Machine Learning Based Spam Mail Identification Approaches. International Journal of Innovative Research in Computer and Communication Engineering, 9(7): 8987-8993.

[23] Jayant Batra, Kirti Bhatia, Rohini Sharma, Shalini Bhadola. (2021). Development and Analysis of SPAM MAIL Identification Model, International Journal of Innovative Research in Science, Engineering and Technology, 10(8): 11528-11535.