IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

An Overview on Machine Learning Based Spam Mail Identification Approaches

Jayant Batra, Kirti Bhatia, Rohini Sharma, Shalini Bhadola

PG Student, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India

Assistant Professor, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India

Assistant Professor, Government College for Women, Rohtak, India

Assistant Professor, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India

ABSTRACT: The increased number of unsolicited emails known as spam has necessitated the progress of increasingly reliable and strong antispam identifiers. Recent machine learning approaches have been successful in detecting and screening spam emails. Anorderly evaluation of certain most widely used machine learning relying email spam screening algorithms is presented. The analysis includes an overview of key ideas, efforts, competence, and the current study tendency in spam screening. The beginning conversation in the study contextual looks at how machine learning techniques are being smeared to the email spam screening processes of major internet service providers including Outlook, Gmail and Yahoo.The overall email spam screening process was discussed, as well as the many initiatives by different researchers to combat spam using machine learning approaches. Our analysis examines the benefits and downsides of existing machine learning algorithms as well as open spam screening research challenges.Deep learning and deep adversarial learning were indicated as future strategies for effectively dealing with the threat of spam emails.

KEYWORDS: Spam e-mail; Machine learning Procedures; Screening

I. INTRODUCTION

Spam, or unwelcome marketable massemails, has become a major issue on the internet in recent years. The spammer is the individual who deliversout the spam mails. This person gathers email IDs from a variety of sources, including websites, chat groups, and viruses. [1]. Spam prohibits users from getting the most out of their time, storage space, and network bandwidth. The giganticquantity of spam mails streaming across communication networks has a negative effect on email servers' storage, communication capacity, processingcapacity, and precious time of user [2]. Spam email is becoming a bigger issue every year, accounting for more than 77 percent of all global email traffic. It irritates an individual who obtainunrequested spams.Many users have suffered untold financial losses as a result of internet cheats and other fake practises by spammers who propel emails posing as from trustworthybusinesses in order to encouragepeople to revealconfidentialprivatedetails such as PINs, Bank Authentication Numbers, and credit card figures.

Prominent email sourceslike Gmail, Yahoo, and Outlook have combined diverse machine learning (ML) methodslike Neural Networks (NN) in their spam sieves to efficiently tackle the threat posed by email spams.By evaluating a hugequantity of spam and phishing communications over a hugequantityof machines, these machine learning approaches can study and recognize spam and phishing mails. These spam sieves examine trash emails using a standard criteria, thanks to machine learning's ability to adapt to changing situations.As they attempt to scan for spam, activity, they build freshdirectionsestablished on what they've learned. Google's machine learning model has progressed to such efficiency that it can now recognize and sieve spam and phishing emails with more than 99 %precision rate.The inference is that only one message out of onehundred manages to get past their spam sieve. According to Google statistics, most of the emails received by Gmail are unwanted.Google's recognition prototypes now include technologies like Google Safe Surfing, which may recognize websites with harmful URLs.

II. RELATED WORK

The international scientific community is becoming increasingly interested in email spam screening. Analogous assessments that have been be stowed in the same research topic are presented in this section. This strategy is used to pronounce the problems that need to be spoken as well as to emphasize the contrasts between our present study and the previous one. In [3,] the authors conducted a quick survey to see whether information screening and recovery

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

processmay be used to hypothesize spam mail discovery in a reasonable, hypothetically justified way, to enable the implementation of spam screening techniques that are effective. However, the poll did not go into detail about the machine learning methods, simulation tools, publicly available datasets, or the spam mail situation's architecture.It also falls short of supplying the metrics that have been used in earlier studies to assess other proposed approaches. The authors of [4] discussed the many approaches employed to sieve out unwanted spam mails.In addition, the article categorizes email spam into separate hierarchical files and automates the procedures required to respond to each email message. The ML approaches, email spam design, relative study of earlier procedures, and the experimentalsetup were all not discussed, which is one of the review article's weaknesses. The authors of [5] analyzed formerlyproposed spam screeningprocedures, with anemphasis on the proposed framework' efficiency.The review's major goal is to look into the connections amid email spam screening and other spam screening systems used in communiqué and stowagemeans. The research also looked into the classification of email spams, as well as the person's details needs and the purpose of the spam sieve as part of a big and intricate information structure.Definite crucial aspects of spam sieves, however, were not taken into account in the review. These include the system's design, the simulation setting, and a relativestudy of the enactment of the sieves under consideration.The authors of [6] discussed the research difficulties surrounding email spam, how it impacts users, and how users and providers might mitigate its consequences. The study also lists the lawful, financial, and procedural approaches that have been utilized to combat email spam.They noted that, relying on methodical metrics, subject examinationsieves have been widely utilized and have been shown to have a decent proportion of correctness andexactness; so, the evaluation concentrated more on them, outlining how they function.The article outlined the structure and operation of a number of machine learning algorithms that are used to sieve email spam. [7] presented a trivialreview on E-mail image spam screening algorithms.The research focused on email antispam screening technologies that were utilized to transition from text-centered to image-centeredapproaches. Spam and the spam sieves designed to combat it have sparked a rise of inventiveness and creativity.The ML approaches, simulation software, emaildataset, and the design of email spam screening strategies were not covered in the study. The authors of [8] provided a comprehensive overview of certain most widely used content-centered e-mail spam screening approaches.The ML approaches for spam screening were the subject of the paper. They looked at the key principles, efforts, efficacy, and spam screening trends.They reviewed the principles of e-mail spam screening, the varyingtypeof spam, advertisers' strategies to get through spam sieves at ESPs, and common MLapproaches for combating spams.

III. SPAM MAIL SCREENING PROCESS

The header and the body are the essential parts of an email message. The header is the section of the email that contains general information about the content. The subject, sender, and receiver are all included. The email's body is its crucial component.It can incorporate data that hasn't been pre-defined. Web pages, audio, video, analogue data, photos, files, and HTML syntax are all examples.The email header contains information like the sender and the recipient's addresses, and the timeframe that show when the message was transmitted through midway servers to the Message Transport Agents, who act as a sort of mail office. A “From” is usually the first word in the header line and When it transfers from one server to another via an intermediary server, it undergoes some changes.The user may see the email's path and the time it takes each server to process it by looking at the headers. Before the classifier can use the provided data for screening, it must go through certain processing.

The steps that must be followed while mining data from an email message are divided into the subsequent categories:

- Pre-processing: When an incoming message is received, this is the first stage that is performed. Tokenization is the first step in this process.
- Tokenization: This is a method for removing words from the body of an email. It also breaks down a message into its constituent elements. It separates the email into a series of representing symbols known as tokens.The authors of [9] stressed that these demonstrative signs are taken from the email text, header, and topic.The authors claimed in [16] that by substituting information with individualI Dsigns, all of the features and verses in the email will be removed, with the exception of the meaning.

- Feature selection: The feature selection stage comes after the pre-processing stage. Feature selection is a type of 3-Dexposure reduction that effectually epitomizes intriguing email message portions as a compacted feature vector. When the message size is huge and aabbreviated trait demonstration is required to build text or image coordinating quick [10], the technique is advantageous. Stemming, noise elimination, and stop word elimination are all part of the feature selection process. Fig. 1 shows a complete process of Spam Email filtration.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

IV. DIVERSE GROUPS OF SPAM SCREENING METHODS

A. *Techniques for Content-Based Screening:*The ML algorithms like Nave Bayesian classification, Support Vector Machine (SVM), K-NN, and NN are commonly employed to construct automatedscreeninginstructions and to categorize emails using content based screening.This schemeinvestigates terms, the existence, and dispersals of verses and idioms in email matter, and then employs developed guidelines to sieve incoming spam emails [11].

B. *Spam Screening Method Using Cases:* One of the most often used spam screening approaches is case-based or sample-based screening. First, using a collection approach, all emails, both ham and spam, are mined from each person's email.Following that, pre-processing procedures are performed to convert the email utilising the interface of client, feature mining and assortment, email data aggregation, and process evaluation. The information is then divided into two vector groups.Finally, a MLapproach is utilised to train and evaluate datasets in order to determine if incoming emails are spam or not[11].

C. *Spam Screening Techniques Using Heuristics or Rules:* This method compares a large figure of patterns, mostly are frequentwords, to a selected message using previously developed rules or heuristics. Several comparable patterns boost a message's score.It, on the other hand, subtracts points if any of the forms do not match. Any communication with a total more than a certain inception is considered spam; otherwise, it is considered genuine.Although certain ranking rules do not alter over time, others must be reorganizedfrequently to deal excellently with the threat of advertisers who constantly sendfresh spam mailswhichmayeffortlessly slip through email sieves [11]. SpamAssassin is a nice instance of an instruction cantered spam sieve[12].

D. *Former Likeness Based Spam ScreeningMethod:* This method classifies arriving emails relyingupon their similarities to warehousedinstances using reminiscence centred, or case centred, ML approaches. The email's properties are utilised to generate a 3D space vector that is then employed to draw freshcases as points. Following that, the freshcases are allocated to the best class of its K-nearest training cases [13]. For spam email screening, this method employs the k-nearest neighbour (kNN) algorithm.

E. *Algorithm for Adaptive Spam Screening:* Spam is detected and sieved using this method, which divides spam into several types. It separates an email dataset into different classes, each with its own distinctive text. Each incoming email is compared to every class, and a portion of equivalence is calculated to determine the most likely class to which the email fits [14].

Numerous academics and researchers have developed many email spam categorization procedures that have been effectively utilised to group data.Probabilistic, decision tree, artificial immune system [15], (SVM) [16], artificial neural networks (ANN) [17], and case-centred strategy [18] are some of these methods.It has been demonstrated in the literature that these classification approaches may be used for spam mail screening by employing a content-centredscreening methodology that recognises specific traits.The frequency with which these characteristics exist in emails determines the probability for each feature in the email, which is then compared to a threshold value.Spam is defined as email messages that surpass a certain threshold value [19]. The artificial neural network is a non-linear framework that attempts to mimic the activities of biological NN.

V. MACHINE LEARNING BASED E-MAIL CLASSIFICATION

The MLapproaches are now being employed to classify spam mail. These algorithms are designed to differentiateamong spam and HAM mails. This is accomplished by the use of an automatic and adaptive technique by machine learning techniques.ML approaches have the ability to extract knowledge from a collection of mails, and then utilize that info to categorize fresh mails that it has justobtained, instead of relying on instructions which are sensitive to the constantly changing features of spam communications.In this unit, we'll look at some significantML approaches for spam identification. Fig. 2 represents different ML approaches for SPAM identification.

A. *Clustering procedure*: It is the process of categorizing a set of patterns into connected categories. It is a technique for categorizing items or case examinations into groups that are comparatively analogous. Thesepractices have recently attracted the interest of many academics and scholars, and they have been utilized in a variety of domains. On e-mail spam corpus with true labels, clustering methods, which are unsupervised learning tools, are applied. A large number of clustering methods can categories e-mail spam corpus into either legitimate mail or unsolicited mail clusters if adequate representations are available. The authors demonstrated this in their research on e-mail spam clustering in [23]. The results were particularly interesting because their procedure outperformed existing recent semi-supervised techniques, suggesting that clustering can be a powerful technique for screening spam e-mails. It groups items or thoughts together in such a way that objects in the same group are more similar than those in other groups. Density-based clustering and K-NN are two kinds of clustering approaches that have been utilized

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

for spam categorization. kNN is a distribution-agnostic algorithm that does not assume the data is derived from a particular likelihood dissemination [24]. The kNN is termed as a lazy learner since It does not do generalisation using the training data points.Hence, no evident training phase exists, and if it does, it is quite limited. The conclusion is that the training stage of the method is quite quick. Due to the lack of universality, kNN must keep all of the training data. During the testing stage, the entire training data set is required since judgments are reliedupon the entire training data. There is a glaring contradiction here in that there is no substantial training period, but there is a lengthy testing period. Both time and memory have an overhead cost. In the worst-case scenario, additional time may be required. To archive all of the training data neighbors, more RAM is required.

B. *Support Vector Machines (SVM):* Over time, it has shown to be a potent and proficientadvanced categorizationapproaches for combating email spam [25]. They're supervised learning prototypes that examine data and find forms that can be used to categorize and investigate relationships amid variables of concern.SVM aapproachesare extremely effective in detecting patterns and categorizing them into specific classes or groups. They are simple to train, and some researchers claim that they outperform many standard email spam categorization systems [26].Because SVM uses data from the email corpus during training, this is the case. However, because to the computational difficulties of the processed data, the power and usefulness of SVM for high dimension data diminishes over time [27].It is a worthy classifier, according to [28], because of its scant data presentation and noblememory and correctness values. The classification accuracy of SVM is very high. Furthermore, SVM is a well-known case of "kernel techniques," which is an important area of ML approaches.

C. *Decision tree(DT):* It's yet another ML method that's been used to magnificently screen spam emails. During the training of datasets, (DT need very little work from consumers.DT is in charge of the hypothesis testing and feature engineering for the email corpus data training. The relationships between parameters have no bearing on a tree's performance. One of the most useful features of a DT is its ability to allocatedefinite values to problems, conclusions, and decision outcomes [29].This reduces the ambiguity in conclusion building. One more significant benefit of the DT over other MLapproaches is that it opens all possible possibilities and pursues each one to its conclusion in one perspective, allowing for clear estimation among the tree's various nodes. Despite its many benefits, the Decision tree does have some disadvantages, including the fact that managing tree development can be problematic without proper pruning. The DTs are a type of nonparametric ML method that is extremely versatile but also susceptible to overfitting of training data [29].As a result, they are somewhat weak classifiers and their classification accuracy is limited.NBTree based classification [30], C4.5/J48 DTProcedure [31], and LMT DT [30] are three different forms of DTs that have been used in spam mail screening. Starting at the root and working your way up the tree until you reach a leaf node that offers the categorization outcome, a DT can be used to provide a solution to a classification challenge.The approach of decision tree learning has been used in spam screening. The objective is to build a DTmodel and train it so that it can predict the value of a goal variable reliedupon a set of input variables. Some of the input variables are communicated with by the relevant inner node [32].The generated email corpus has the largest info gain and so contamination (both spam and ham) of the sample is decreased by partitioning the email dataset according to least entropy.The decision tree approach can be used to test the dataset after the tree has been created from the training email dataset. In order to get to a leaf node, the email dataset being evaluated goes through certain processing in the tree using some established criteria.The tested data is then given the label from the leaf node.

D. *Naïve Bayes*: It is also a fantastic ML method that has been used in email spam screening is this one. A Naive Bayes (NB) classifier applies Bayes' proposition to the context categorizationof all emails, assuming that the words in the email are unrelated to one another [33].When equated to conditional prototypes like logistic regression [28], NB is preferable for spam mail screening due of its easiness, simplicity of enactment, and speedy merging. It only requires a small amount of training data. It's incredibly adaptable.Increases in the number of predictors and discrete units of information do not generate bottlenecks [28]. NB may be used to solve both two-class and multi-class classification issues.It can be applied on forecast events that are subject to or include probability disparity. They have the ability to manage both continuous and discrete data successfully. Irrelevant features have no effect on NB algorithms. The Nave Bayes algorithm is widely used in commercial and open-source spam sieves [34]. In addition to the benefits described above, NB requires minimum training time and can detect and sieve email spam quickly. The prior collection of non-spam and spam texts can provide training for NB sieves [28]. It maintains track of the changes that occur in each word in legitimate and illegitimate messages, as well as in both.NB may be used to detect spam messages in a variety of datasets with varying features and attributes [28]. The Bayesian classification technique exemplifies both supervised learning and statistical classification techniques. It functions as a foundational probabilistic model that allows us to exploit ambiguity in the model in an ethical manner by affecting

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

the results' probabilities. It's utilized to solve problems that are both analytical and predictive [35].The classification includes practical learning techniques, as well as the ability to combine prior information and investigational data. Bayesian Classification provides a useful perspective for studying and evaluating a variety of learning techniques. It is resilient to noise in input data and computes accurate likelihoods for postulation.A Naive Bayes classifier is a simple probabilistic classifier based on the Bayes theorem and based on reasonable conventions that are self-contained.

E. *Neural networks:*ANNare interconnected clusters of modestcomputationmodules that interconnect with one another via a large figure of weighted networks. Every unit takes input from nearby units as well as outside sources and computes an output that is sent to other units.There is also a channel for refinement the weights of the networks. Neural networks are a powerful approach for handling any classification-related machine-learning problem [35]. They are growing as a prominent tool in the MLscholar's toolkit due to their resourcefulness.However, as one might expect, NN are not often used in the recognition of spam mail. Nearly all modern spam sieves employ naive Bayes classifiers as an alternative. When the term "analogue neural network" is used, it usually refers to one of two types of neural networks.The perceptron and the multilayer perceptron are the two types of perceptron. Figure 3 shows a neural network approach to detecting SPAM mail.

F. *Firefly algorithm:* In[36], authors presented the firefly algorithm (FA), which is a populace centered metaheuristic procedure. He was inspired by the dazzling behavior of fireflies.To direct the search, the algorithm uses population physiognomies to conserve and increase several candidate solutions [37]. The programme was created based on research into the idea of communiquéamid fireflies as they prepare to couple and are instantaneouslyshowing to threat.Fireflies use their brilliant quality to communicate with one another [38]. Each of the world's approximately 2000 firefly species employs a unique dazzling style.The firefly usually produces a small spark with a specific pattern depending on what they are doing. The light is produced by living creatures' biochemical light synthesis.Based on the light's shape, the proper companion will respond by either copying the same form or responding with an exact form. The intensity of light, on the other hand, decreases as distance increases.As a result, a firefly's glittering light attracts the attention of other fireflies within the flash's visual range.

G. *Rough set classifiers:*The technique focuses on the decomposition of categorization of vague, abstruse, or limitedinfo expressedas experience data. Rough set concept is a relatively new mathematical method for dealing with fuzziness.Rough Set is based on the premise that everything in the universe has some knowledge associated with it. RS is a geometricmechanism that focuses on the ambiguity of a situation [39].It is in agreement with the idea that any approximateprototype can be predictable from below and above via a natural relationship.The necessity to find redundancy and relationships between features is one of the main characteristics of the RS philosophy [40]. Rough Set theory is used in spam screening because it provides fast and efficient techniques for extracting hidden patterns in data.It also has the ability to quickly recognize associations that are difficult to find using traditional statistical techniques. Furthermore, it allows both quantitative and qualitative data to be used.It is proficient of estimating the minimum data sets required for job grouping. Fig. 4demonstrates the email screeningprocedure of the RSmethod from the mailbox of a user.

H. *Ensemble classifiers:* It is a new tacticwhere a collection of diverse classifiers is trained and combined to increase the categorization accurateness of the entire system on the same issue, in this instance spam screening.They are a type of ML approaches that operate together to enhance the overall categorizationperformance of the scheme. The authors of [41] called for the assemblage of many sieves as a fascinating strategy to efficiently address spam, which today arrives in a variety of formats.Bagging and boosting [42] are the most extensively used ensemble classifiers. These techniques use subsets of the entire data set to train classifier instances. Bagging aggregates the outputs of trained classifiers on a sample of the data set selected from a bigger sample.The most often used boosting method is AdaBoost. [43] made the suggestion. Even when the weak learners' performance is subpar, AdaBoost can produce a good result. Boosting is presently used in the fields of classification, regression, and facial recognition, among other things.

I. *Random forests (RF):*It is is an ensemble learning strategy and regression method that can be used to solve problems involving data classification into groups [44]. Decision trees are used by the algorithm to make predictions.The programmer writer creates several decision trees during the training step. Following that, these decision trees are used to predict the set; this is performed by considering the designatedsets of every separate tree, and the set with the most votes is chosen as the outcome.The RF technology is rising in reputationnow, and it is

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

being used in a variety of industries. It provides a quick way to calculate the estimated value of misplaced data while maintaining correctness in instances where a large section of the data is missing.The user can grow as many trees as they like using RF.

J. *Deep learning (DL) procedures:* It is a rapidly growing field that uses machine learning and artificial intelligence to learn characteristics straight from data utilizing numerous nonlinear processing layers. In email spam classification, DLprototypes can attain extraordinaryprecision.Deng and Yu [45] addressed several deep learning techniques, including how to classify them as supervised, unsupervised, or HDN based on their topologies and applications.There are three types of layers in CNNs. In CNNs, the stacking up of numerous layers allows for automatic learning of highly discriminative feature descriptions without the requirement for manual features.A CNN differs from a standard BPN in that a BPN works with isolated hand-crafted image data, but a CNN works with an email message to extract important, critical qualitiesfor categorization.

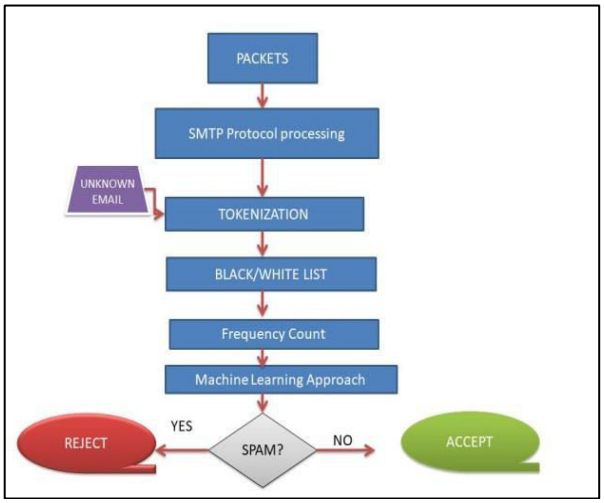

Fig.1.Spam Mail Filtration Process

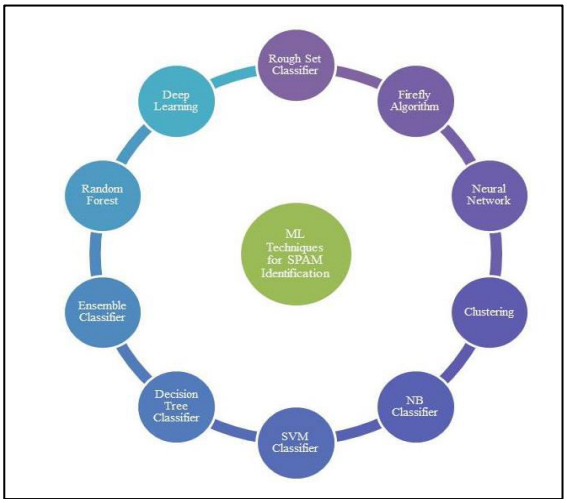

Fig. 2. Different Machine Learning Approaches for SPAM mail Identification

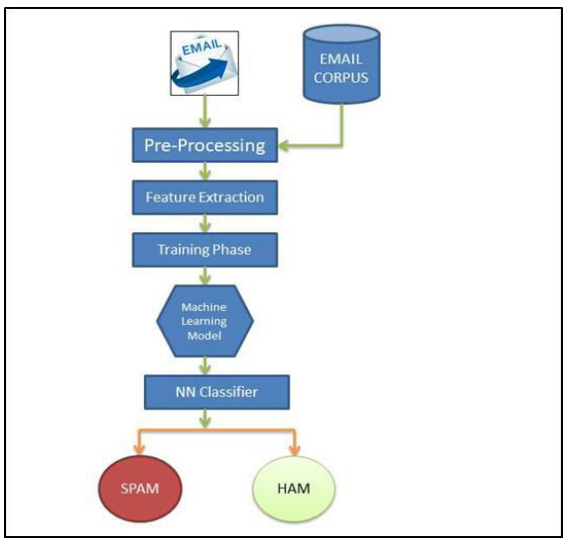

Fig. 3.Neural Network Approach for SPAM Mail Identification

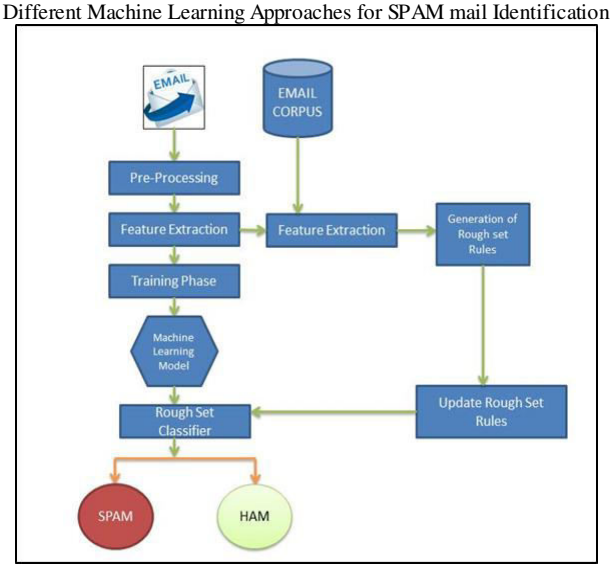

Fig 4.Rough Set classifier based SPAM Mail identification

VI. CONCLUSION AND FUTURE WORK

In this article, we looked at ML techniques and how they can be applied to spam screening. Ananalysis of current advanced procedures for classifying communications as spam or ham may be found here.It was highlighted how various academicians attempted to solve the issue of spam using machine learning classifiers. Spam communications

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

have evolved over time to bypass sieves, which has been studied.The basic design of an email spam sieve, as well as the processes involved in spam email screening, was investigated. The study looked at certainopenly accessible datasets and enactment measures which may be employed to assess the efficacy of spam sieves.The difficulties of MLprocedures in effectively dealing with the spam threat were highlighted, and relative assessments of ML techniques accessible in the literature were conducted. We have discovered several unsolved research issues with spam sieves. Overall, the amount and number of material we examined indicate that tremendous growth has been done and will continue to be rendered in this sector. Following the discussion of the outstanding issues in spam screening, more research to enhance the efficiency of spam sieves is required. As a result, academics and businessconsultants investigating ML approaches for actual spam screening will continue to be engaged in the development of spam sieves. We expect that this study will serve as a springboard for high-quality study in spam screening employing ML, deep learning, and deep adversarial learning procedures by research students.

REFERENCES

1. M. Awad, M. Foqaha, 'Email spam classification using hybrid approach of RBF neural network and particle swarm optimization', International Journal of Network Securityand its Applications, Vol. 8, Issue4,2016.
2. D.M. Fonseca, O.H. Fazzion, E. Cunha, I. Las-Casas, P.D. Guedes, W. Meira,M. Chaves, 'Measuring characterizing, and avoiding spam traffic costs', IEEE Int.Comp. 99 2016.
3. C.P. Lueg, 'From spam screening to information retrieval and back: seeking conceptual foundations for spam screening', Proceedings of the Association for Information Science and Technology, Vol. 42, Issue1, 2005.
4. X.L. Wang, 'Learning to classify email: a survey', in: 2005 International Conference on Machine Learning and Cybernetics. Vol. 9, pp. 5716-5719, IEEE, Aug 2005.
5. G.V. Cormack, 'Email spam screening: a systematic review', Found. Trends Inf. Retr, Vol.1, Issue 4, pp. 335–455, 2008.
6. E.P. Sanz, J.M.G. Hidalgo, J.C.C. Perez, 'Email spam screening', Advanced Computing, Vol. 74,pp. 45–114, 2008.
7. S. Dhanaraj, V. Karthikeyani, 'A study on e-mail image spam screening techniques, in International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013.
8. A. Bhowmick, S.M. Hazarika, 'Machine Learning for E-Mail Spam Screening: Review, Techniques and Trends', arXiv:1606.01042v1 [cs.LG], pp. 1–27, 2016.
9. T. Subramaniam, H.A. Jalab, A.Y. Taqa, 'Overview of textual anti-spam screeningtechniques', International Journal of Physical Sciences, Vol.5, Issue 12, pp.1869–1882, 2010.
10. M.F. Porter, An algorithm for suffix stripping, Program: The Electronic Journal of Information Systems, Vol. 14, Issue 3, pp. 130–137, 1980.
11. V. Christina, S. Karpagavalli, G. Suganya, 'Email spam screening using supervised machine learning techniques', International Journal of Computer Science Engineering, Vol. 02, Issue 09, pp. 3126–3129, 2010.
12. J.R. Mendez, F. Díaz, E.L. Iglesias, J.M. Corchado, 'A comparative performance study of feature selection methods for the anti-spam screening domain, in: Advances in Data Mining. Applications in Medicine', Web Mining, Marketing, Image and Signal Mining, Springer Berlin Heidelberg, pp. 106–120, 2006.
13. G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, 'Stacking classifiers for anti-spam screening of E-mail', in: Empirical Methods in Natural Language Processing, pp. 44–50, 2001.
14. L. Pelletier, J. Almhana, V. Choulakian, 'Adaptive screening of spam', in: SecondAnnual Conference on Communication Networks and Services Research(CNSR'04), 2004.
15. E.M. Bahgat, S. Rady, W. Gad, 'An e-mail screening approach using classification techniques', in: The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Springer International Publishing, BeniSuef, Egypt, 2016, pp. 321–331.
16. N. Bouguila, O. Amayri, 'A discrete mixture-based kernel for SVMs: application to spam and image categorization', Information Processing & Management, Vol. 45, Issue6,pp. 631–642, 2009.
17. Y. Cao, X. Liao, Y. Li, 'An e-mail screening approach using neural network', in:International Symposium on Neural Networks, Springer Berlin Heidelberg, pp. 688–694, 2004.
18. F. Fdez-Riverola, E.L. Iglesias, F. Diaz, J.R. Mendez, J.M. Corchado, 'SpamHunting: an instance-based reasoning system for spam labelling and screening', Decision Support System, Vol. 43, Issue3,pp. 722–736, 2007.
19. S. Mason, New Law Designed to Limit Amount of Spam in E-Mail, 2003.http://www.wral.com/technolog.
20. I. Stuart, S.H. Cha, C. Tappert, 'A neural network classifier for junk e-mail, in: Document Analysis Systems VI', Springer Berlin Heidelberg, pp. 442–450, 2004.
21. J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Elsevier, 2011.
22. T.S. Guzella, W.M. Caminhas, A review of machine learning approaches to spamscreening, Expert Systems with Applications, Vol. 36, Issue7,pp.10206–10222, 2009.
23. J.S. Whissell, C.L.A. Clarke, 'Clustering for semi-supervised spam screening', in: Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS '11), 2011, pp. 125–134.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907200 |

24. T. Saravanan, 'A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm', Retrieved on August 8, 2017 from, 2010.
25. Z.S. Torabi, M.H. Nadimi-Shahraki, A. Nabiollahi, 'Efficient support vector machines for spam detection: a survey', IJCSIS, Vol. 13, Issue 1,pp. 11–28, 2015.
26. B. Sch€olkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond, MIT press, 2002.
27. B. Yu, Z. Xu, A comparative study for content-based dynamic spam classification using four Machine Learning algorithms, Knowledge Based System, Vol. 21, Issue4,pp. 355–362, 2008.
28. P. Chhabra, R.hWadhvani, S.Shukla, 'Spam screening using support vector machine', Special Issue of IJCCT, Vol. 1, Issue 2, International Conference [ACCTA-2010], Pp. 166-171, 2010.
29. Jason Brownlee, Master Machine Learning Algorithms, Discover How They Work and Implement Them From Scrath, 2019.
30. S. Chakraborty, B. Mondal, 'Spam mail screening technique using different decision tree classifiers through data mining approach - a comparative performance analysis', International Journal of Computer Application, Vol. 47, Issue 16,pp.0975 – 888, 2012.
31. K. Masud, M.R. Rashedur, Decision tree and naïve Bayes algorithm for classification and generation of actionable knowledge for direct marketing, Journal of Software Engineering and Applications, Vol. 6, pp. 196–206, 2013.
32. G. Holmes, G. Pfahringer, B. Kirkby, R. Frank, E.M. Hall, Multiclass Alternating Decision Trees, ECML, 161–172, 2002.
33. I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, C.D. Spyropoulos, 'An evaluation of naive bayesian anti-spam screening', in: Proceedings of 11th European Conference on Machine Learning (ECML 2000), Barcelona, pp. 9–17, 2001.
34. I. Androutsopoulos, G. Paliouras, E. Michelakis, Learning to Sieve Unsolicited Commercial E-Mail. Tech. Rep., National Centre for Scientific Research Demokritos, Athens, Greece, 2011.
35. G. Bandana, Design and Development of Naïve Bayes Classifier, North Dakota State University of Agriculture and Applied Science, Graduate Faculty of Computer Science, 2013. Master thesis A. Edstrom, Detecting Spam with Artificial Neural Networks, Retrieved on August 10, 2017.
36. X.S. Yang, Firefly algorithms for multimodal optimisation, Proc. 5th symposium on stochastic algorithms, foundations and applications, in: O. Watanabe, T. Zeugmann (Eds.), Lecture Notes in Computer Science 5792, 2009, pp. 169–178.
37. J. Dugonik, I. Fister, Multi-population firefly algorithm, in: Proc. Of the 1stStudent Computer Science Research Conference, Ljubljana, Slovenia, pp. 19–23, 2014.
38. W.A. Khan, N.N. Hamadneh, S.L. Tilahun, J.M. Ngnotchouye,' A Review and Comparative Study of Firefly Algorithm and its Modified Versions', Intech Publishing House, pp. 281–313, 2016.
39. S.S. Roy, V.M. Viswanatham, P.V. Krishna, N. Saraf, A. Gupta, R. Mishra, 'Applicability of rough set technique for data investigation and optimization of intrusion detection system', in: Quality, Reliability, Security and Robustness in Heterogeneous Networks, Springer Berlin Heidelberg, pp. 479–484, 2013.
40. N. Perez-Díaz, D. Ruano-Ordas, F. Fdez-Riverola, J.R. Mendez, 'Rough sets for spam screening: selecting appropriate decision rules for boundary classification', Appl. Soft Comput, Vol. 13,Isuue8, pp. 1–8, 2012.
41. P.H.C. Guerra, D. Guedes, J.W. Meira, C. Hoepers, M.H.P.C. Chaves, K. Steding- Jessen, Exploring the spam arms race to characterize spam evolution, in: Proceedings of the 7th Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), Redmond, WA, 2010, July.
42. L. Breiman, Bagging predictors, Mach. Learn, Vol. 24, Issue 2,pp.123–14, 1996.
43. Y. Freund, R.E. Schapire, A Decision - theoretic generalization of on - line learning and an application to boosting, JCSS 55, pp. 119–139, 1997.
44. A.A. Akinyelu, A.O. Adewumi, Classification of phishing email using random forest machine learning technique, J. Appl. Math. 6 (2016). Article ID 425731, Retrieved on July 12, 2017.
45. L. Deng, D. Yu, Deep Learning: Methods and Applications, Now publishers, Boston,2014.

BIOGRAPHY

Jayant Batrais a M.Tech. Student of Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India.He has done B.tech. in computer science from Vaish College, Rohtak. His area of Interest is Deep machine learning.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details