Detection of Severity-based Email SPAM Messages using Adaptive Threshold Driven Clustering

I V S Venugopal[1], D Lalitha Bhaskari[2], M N Seetaramanath[3]
Department of IT, G V P College of Engineering (A), Andhra Pradesh, 530048, India[1, 3]
Department of CS&SE, AUCE (A), Andhra Pradesh, Visakhapatnam, India[2]

***Abstract*—The classification of emails is one crucial part of the email filtering process, as emails have become one of the key methods of communication. The process for identifying safe or unsafe emails is complex due to the diversified use of the language. Nonetheless, most of the parallel research outcomes have demonstrated significant benchmarks in identifying email spam. However, the standard processes can only identify the emails as spam or ham. Henceforth, a detailed classification of the emails has not been achieved. Thus, this work proposes a novel method for the identification of the emails into various classes using the proposed deep clustering process with the help of the ranking of words into severity. The proposed work demonstrates nearly 99.4% accuracy in detecting and classifying the emails into a total of five classes.**

***Keywords*—BoW collection; web crawler; email text extraction; subsetting method; email class detection; ranking method**

I. Introduction

The increase of email communication and people getting higher focus on the email-based communications have opened the door for hackers to target more and more people to be trapped using spam emails.

Email spam prevention remains a difficulty due to attackers using novel ways that evade current spam filters, as noted by H. V. Bathala et al. [1]. An email filter that can detect zero-day attacks is essential. The standard approach offers a method that not only examines the email's body content but also the phishing URLs and spam pictures attached to it. Machine Learning techniques are used to categorise the emails, and a systematic procedure is provided to detect spams. The existing machine learning models are selected using the lazyPredict library. When tested with standard data sets, these smart filters are effective at detecting spam and guarding against zero-day threats. For URL phishing detection, the Stacking classifier performs better with an accuracy of 0.97. In contrast, the perceptron classifier, which has an accuracy of 0.97, is the best at spotting spam email. Other algorithms' results are also included.

M. Hina and colleagues [2] stress the relevance of the detection procedure in their study since email has been used for a long time as a safe and trustworthy method for communication. Email's importance has grown because of the proliferation of quick and secure communication methods. Email management has become more difficult due to the tremendous growth of email data. Until now, emails have been categorised and sorted by sender, size, and date of origin. But there is a need to identify and categorise emails based on their content. There have been a variety of methods used in the past to classify emails as either spam or non-spam based on their content. A multi-label email categorization system is proposed in the standard approach. Forensic studies of huge email data have been suggested to use an efficient categorization system (e.g., a disc image of an email server). An investigator using this technique would have an advantage when looking into crimes using email. It was shown that Logistic Regression outperforms Naive Bayes, Stochastic Gradient Descent, Random Forest and Support Vector Machines in a comparison of machine learning algorithms. Using benchmark data sets, logistic regression was shown to be the most accurate, with a 91.9 percent accuracy rate for bi-gram features. The similar recommendations can also be observed in the work by Naser et al. [21].

M. K. Islam et al. [3] have shown that email communication is crucial in many aspects of daily life, notably in the workplace. Spam email filtering is critical, especially considering its pervasiveness. Sending massive amounts of undesired communications through electronic means known as spam email, or junk email, constitutes this kind of spam. Due to hazardous frauds or malware hosting sites or viruses attached to the message, most spam emails are not only bothersome but also destructive. In standard measures, the researchers have uncovered the characteristics that set spam apart from other types of email. Including the pooled dataset, the researchers have used four machine learning models and two deep learning models. In addition, the researchers have searched the spam email collection for crucial terms that appear regularly. The researchers have may use this information to identify spam emails for the safety of the employees and the community.

However, the standard mechanism for detection of the spam emails is highly influenced by the use or choice of languages. Hence, many of the standard emails with valid text and information are also often marked as spams causing confusions. Thus, this work decides to solve this problem with deeper classification of the email using proposed machine learning method.

II. Foundational Strategy for Email Classification

After setting the context in the previous section of this work, in this section of the work, the foundational strategy for email classification is discussed.

Assuming that the total text or the collection of the email texts is LT[] and each email set is T[]. Thus, for n number of emails in the total collection can be formulated as,

$$LT[] = < T[0], T[1], T[2], ...T[n] > \quad (1)$$

Further, each and every email corpus is collection of sender details, receiver details, subject and the actual email text. This can be formulated as,

$$T[] = < Sender, \text{Re}ceiver, Subject, Text > \quad (2)$$

Furthermore, the set of words, which will decide the nature of the email, or popularly called bag or words, denoted as BOW[], also must be populated. Assuming the number of elements in BOW[] are m and each word is denoted as W_X, thus, this can be formulated as,

$$BOW[] = < W_1, W_2, ...W_m > \quad (3)$$

As per the foundational strategy, the frequency of any word from the BOW[] set must be cross checked in the in the email text and the frequency of the word must be recorded. This can be formulated as,

$$Fq\{W_X\} = \frac{\phi\{W_X \prec T[i]\{Text\}\}}{\phi\{T[i]\}} \quad (4)$$

Where, W_X is the frequency calculated using the function Fq{} and ϕ is the function for extracting the count of the searching word.

Furthermore, the same frequency of the word, $Fq\{W'_X\}$, must be identified in all the text available in the dataset, that is, LT[].

This can be formulated as,

$$Fq\{W'_X\} = \frac{\phi\{W_X \prec LT[]\{T[i].Text\}\}}{\phi\{LT[]\}} \quad (5)$$

During the final phase of the analysis, the document frequency is found to be higher than the dataset word frequency, then the email can be identified as spam or in case the document level frequency is less, then the email can be marked as ham. As, formulated below,

$$T[i]\{True \mid False\} : Fq\{W_X\} > Fq\{W'_X\}() \quad (6)$$

Thus, this is the foundational process for identification of spam or ham in the email corpuses.

This understanding of the foundational method will help in critically analyzing the recent research outcomes in the next section.

III. Parallel Recent Research Outcomes

After having the foundational understanding of the traditional methods for detection of the spams, in this section of the work, the parallel recent research outcomes are discussed.

Spam, according to a study by C. Bansal et al. [4], is the most talked about topic on the Internet today. When you transmit spam, it's simple for spammers to do so. Thousands of spam emails flood inboxes. Files, contacts, and other sensitive information are stolen from the devices by spammers. Even with the most cutting-edge equipment, it's still challenging. Here, the researchers have used a computerised neural network to demonstrate how the researchers have may invert the frequency of Term Frequency documents (TFIDF). The confusion matrix, accuracy, and precision are used to compare the outcomes. The researchers have noticed a propensity to utilise Kaggle data sets with a lower mix of spammy emails and actual emails to evaluate the applicability of ANN. TFIDF-based TFIDF ANN yields a positive return of 97.58%, according to the outputs.

Email communication has become one of the most cost-effective and efficient methods for official and corporate users because of the widespread availability of internet connections. Every day, hundreds of millions of spam emails are sent and received. Spam detection is necessary to safeguard the privacy of people or organisations. Handling large datasets in machine learning is time- and space-consuming in Spam detection. Features must be selected to exclude those that aren't significant to reduce the time and space complexity. The suggested aim in this work is to reduce the time complexity, the space complexity, and the accuracy of the feature selection approach. Both global and local optimization strategies are used in the proposed feature selection method, which combines the chi2 select best method with the Tree-based feature selection method. Four distinct classifiers are put to the test in a series of experiments. According to the results of S. Sharma et al. [5], the suggested concept performs well on precision, memory, and accuracy efficiency tests.

According to A. Karim et al. [6,] an intelligent and automated anti-spam framework is required due to the explosive proliferation of spam email and the inherent destructive dynamic inside such assaults on a variety of social, personal, and commercial activities. A growing number of attacks, including virus propagation and identity theft, as well as sensitive data theft, monetary and reputational harm, are taking place. There are now several solutions that don't take into consideration the wide variety of features available in email. Artificial Intelligence, particularly unsupervised machine learning, is preferred approach. Unsupervised learning is being investigated in this study to see whether it can be used to group spam and ham emails. The researchers have wanted to create an unsupervised framework that relies only on unsupervised methods and a clustering strategy that use various techniques, principally the email body and the subject header, to do this. An entirely new binary dataset of 22,000 spam and ham emails was used for the clustering (reduced from eleven to ten after the feature reduction). In this research, seven of the ten features were developed specifically to reflect important analytical email properties from several angles. With five algorithms tested, OPTICS provided the best clustering with a 0.26 percent greater average effectiveness than DBSCAN, its closest competitor. OPTICS and DBSCAN had a combined accuracy of 75.76 percent on average.

Spam e-mail has several detrimental effects, including increased communications overload and cybercrime, as shown by the study of S. A. A. Ghaleb et al. [7]. Spam email has become a key weapon for assaults such as cross-site scripting, malware infection, phishing, and cross-site request forgery, etc., which is the most dangerous feature of spam email. The effectiveness of earlier methods of discovery has been weakened by adaptive unsolicited spam. Using a series of six distinct forms of the extended Grasshopper Optimization Algorithm (EGOAs), this study presents a novel Spam Detection System (SDS) architecture, which is studied and integrated with a Multilayer Perceptron (MLP) for advanced spam email detection. Neural Network (NN) models are created as a result of combining MLP and EGOAs in this context (EGOAMLPs). In this study, EGOAs are used to train the MLP to distinguish between spam and non-spam emails. SpamBase, SpamAssassin, and the UK-2011 Webspam benchmark datasets are used to test these models. A kind of spam email is shown to be useful in identifying the models' efficacy. A comparison of the accuracy, detection rate, and false alarm rate for the EGOAs-trained MLP model indicated that it outperformed the other optimization strategies in this study.

For business reasons, email is the most common way of official communication. Despite the popularity of alternative forms of communication, email continues to grow in popularity. Automated email management is critical in today's world, as the number of emails continues to rise. More than half of all emails are considered spam. This proves that spams are a waste of time and resources for email users, since they provide no relevant information. Understanding the various spam email categorization strategies and their mechanisms is essential for spammers to carry out their illicit actions through spam emails. The study primarily focuses on the machine learning methods used to classify spam. The research also includes a complete evaluation and analysis of research done on various machine learning methods and email properties utilised in various Machine Learning methodologies. Future scholars may find it interesting to know about the issues that M. RAZA et al. [8] describe in their work on spam categorization.

Every day, new technologies and tools are created and made available to the public. Every day, new software and websites are being developed because of technological advancements. As the quantity of software products grows, so does the number of people who use them. As R. Al-Haddad et al. [9] point out, hackers and bad persons will take advantage of this opportunity to commit fraud, hack, or fool, particularly nave users. Email has long been the preferred method of communication in a variety of settings, including academia, business, and the arts. Because so many businesses depend on email to communicate with consumers and with other businesses, fraudsters and phishers devote more time and resources to sending fraudulent emails. As spam and scam emails grow increasingly common, hackers are constantly tweaking their emails to make them seem more authentic. In this study, four machine learning algorithms are used to distinguish between real and fraudulent emails. The studies make use of classifiers such as Decision Tree, Random Forest, Nave Bayes, and Support Vector Machine. A fresh collection of 11926 emails, 5183 of which are classified as spam and the remainder as regular (ham) emails, is used to test these classification techniques. The findings reveal that SVM performs best when the attained accuracy is more than 98%.

Many organisations and people have found it more convenient to communicate through e-mail. Spammers use this technique to send unsolicited emails to make money, as shown by the study of S. Gibson et al. [10]. Machine learning algorithms that are improved using bio-inspired methodologies will be shown in this article to identify spam emails. A literature study is conducted to investigate the most effective strategies for analysing various datasets to get high-quality findings. On seven separate email datasets, along with feature extraction and pre-processing, substantial research was done to develop machine learning models utilising Nave Bayes, Support Vector Machine, Random Forest, Decision Tree and Multi-Layer Perceptron. Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) were used to improve the classifiers' performance. Overall, Naive Bayes plus Genetic Algorithm outperformed all other methods. Various machine learning and bio-inspired models are also compared to determine which is the most appropriate model.

I. Saha et al. [11] recently shown that under the COVID-19 pandemic, individuals are obliged to adopt a 'work from home' strategy. Nowadays, the Internet serves as an excellent medium for social contact. People's utter reliance on digital platforms puts them up to scammers. It is illegal to employ phishing to get the login credentials of other people on websites like online banking, internet business, e-commerce and even online classrooms and digital markets. Phishers create bogus websites that seem just like the real thing and then send out spam emails to entice unsuspecting victims. When a spam recipient clicks on a link to one of the bogus websites, the phishing scammers steal the user's login credentials. Researchers have developed a wide range of techniques, including blacklists, whitelists, and antivirus software, to identify phishing websites. Defending against cyberattacks is a never-ending challenge for attackers. Phishing websites may now be detected using a data-driven methodology that employs deep learning. Phishing websites are predicted using a feed-forward neural network known as a multilayer perceptron. The data was gathered from Kaggle and consists of 10,000 websites worth of data. There is a total of 10 of them. Accuracy rates for training and testing of the proposed model are 95 percent and 93 percent, respectively.

The rapid rise in the number of internet users has resulted in an increase in email spam, as N. Kumar et al. [12] report. Illegal and unethical practises, such as phishing and fraud, are taking advantage of them. Sending spam emails with dangerous links is a certain way to bring down your system and get your data. Spammers may easily set up a bogus profile and email account, and they target those who aren't aware of the scams. There is a pressing need to identify spam mails that are fraudulent, and this project will do so using machine learning techniques. The work will discuss machine learning algorithms and apply each algorithm to data sets, and then the best algorithm for email spam detection with the best precision and accuracy will be selected.

Spam emails are becoming more widely used in criminal operations such as identity theft, virus propagation, financial loss, and harm to a company's image, among other things, as their effectiveness and diversity expand. These unlawful activities put the privacy of countless individuals and organisations at risk. Numerous studies have attempted to resolve the problem, but none have been successful. The researchers have felt an intelligent and automated approach is the best strategy for dealing with the problems at hand. There have been just a few research on the use of entirely unsupervised frameworks and algorithms to solve the challenge to far. The researchers have plan to provide an anti-spam framework that depends entirely on unsupervised techniques using a multi-algorithm clustering approach to study and examine the possibilities. The first component of the system is examined in detail in this article, focusing on the domain and header information provided in email headers. In this work, a new way of feature reduction employing an ensemble of 'unsupervised' feature selection algorithms was also explored. It was necessary to employ a dataset of 100,000 unique spam and junk email records as the data source. The following are some of the most important findings: I) Of the six clustering algorithms employed, Spectral and K-means performed well, but OPTICS predicted the best clustering with an average of 3.5% greater efficiency than Spectral and K-means, which were confirmed by a variety of validation techniques. Second, the performance of BIRCH, HDBSCAN, and K-modes was insufficient. As shown in the work of A. Karim et al. [13], the suggested feature reduction framework fulfilled its target with great confidence, and the average balanced accuracy for the optimal three techniques is 94.91 percent.

Phishing e-mails, commonly known as spam, including spear phishing or spam-borne malware have become an increasingly significant issue in recent years, prompting the development of effective, intelligent anti-spam email filters, according to A. Karim et al. [14]. For the identification of clever spam emails, the researchers have thought this survey study on artificial intelligence and machine learning may assist design effective countermeasures. Four components of the email's structure that may be exploited for intelligent analysis were explored in the work: There are MTAs (Mail Transfer Agents) in the heads of emails that offer information about the email's origin, destination, and the number of reroutes it has made along the way. In the SMTP Envelope, the originating source and destination domains and users' identifications are included. (C) The first component of SMTP Data, which includes information such as the sender, recipient, date, and topic. (D) The email body and any attachments are included in this section of the SMTP data. Publications describing each approach were selected, reviewed, and summarised based on the number of relevant papers. The work reveals interesting results, difficulties, and research issues. Research on theoretical and empirical aspects of intelligent spam email identification is now possible thanks to this extensive survey.

There is a lot of evidence, according to M. Gupta et al. [15], that the usage of Short Message Service (SMS) on mobile phones has expanded to such an extent that the devices are occasionally inundated with spam SMS. Additionally, spam communications might cause a user to lose confidential information. Spam emails may be effectively filtered using a variety of content-based machine learning algorithms. Text messages may be classified as either spam or ham based on certain stylistic characteristics. Text message spam detection may be considerably affected by the inclusion of well-known terms, phrases, and acronyms. Based on their accuracy, precision, recall, and CAP Curve, the study compares several classification approaches on various datasets gathered from prior research. Traditional machine learning approaches have been compared to deep learning methods.

Email is the most favoured means of transferring information over the Internet. Detecting spam is one of the most daunting tasks for email users. Detection and filtering methods may help with this. Support Vector Machine (SVM) methods may play a critical role in spam identification. Using the KFCM technique, the researchers have suggested the usage of a weighted SVM for spam detection. Various classes have different weights, which is reflected in the weight variables. The increase in weight value reduces the amount of email that is incorrectly classified. In accordance with Vishagini et al., the researchers have assessed the effect of spam detection using SVM, WSVM with KPCM, and WSVM with KFCM. [16].

Hacking and malicious email communications are becoming more severe security issues, according to S. Chawathe et al. [17]. Detection of malicious email by automated or semi-automatic means is a critical weapon in the fight against such email threats. For this aim, the work reports use fuzzy rules to categorise email. Using a real dataset, researchers test the performance of a classifier based on fuzzy rules against that of classifiers using crisp rules and decision trees. The usability and editability of the classifiers generated by these approaches are also examined.

If you've ever received an e-mail that you didn't want, you've received spam. Spam email filters are becoming more and more necessary for email users because of the ever-increasing number of spam emails they must deal with. With ever-increasing email volumes, spam classifiers are becoming more ill-equipped to manage them and to identify and detect new spam emails with high performance. It's difficult to classify spam because of the large number of characteristics. An essential part of keyword content categorization is selecting features that are among the most common and successful approaches for reducing feature complexity. As a result, any unnecessary or redundant features that would slow down the system will be removed. Meta-heuristic optimization is the process of selecting the best answer from a set of feasible alternatives while keeping in mind the study's primary goal: performance. Other issues include a lack of clarity in regard to the impact of optimization feature selection on prior work's prominent classifier algorithms, such as K-nearest Neighbor, Nave Bayesian, and Support Vector Machine. So, the goal of this study is to increase feature selection accuracy by using a hybrid Water Cycle and Simulated Annealing approach to optimise findings and assess the suggested Spam Detection method. For this study, the researchers followed a five-step approach to conducting their research. The suggested

spam categorization was put to the test using cross-validation on seven different datasets. Meta-heuristic water cycle feature selection (WCFS) was used as a feature selection method in conjunction with Simulated Annealing, as shown by the findings. The hybridization interleaved hybridization surpassed other feature selection algorithms, such as Harmony Search, Genetic Algorithm, and Particle Swarm, with an accuracy of 96.3 percent. On the other hand, the SVM outperformed other classifier algorithms with an f-measurement of 96.3 percent. As indicated by G. Al-Rawashdeh et al. [18], the number of features using interleaved water cycles and Simulated Annealing has lowered by more than half.

In both personal and professional settings, email is a go-to method for exchanging information. Even though electronic communications, mobile apps, and social networks have become more common, e-mails have remained a vital form of communication. There are several reasons why automated email management is necessary, including spam classification, phishing classification, and multifilter categorization, among others, as the number of vital e-mails grows. All articles on email classification from 2006 to 2016 were analysed using the methodological decision analysis in five areas, including e-mail classification application areas, data sets used for each area of classification, feature space used for each application area, and e-mail classification techniques and performance measures [19].

According to research by W. Z. Khan et al. [20], the issue of email spam has expanded dramatically in the last several years. It's not only a hassle for users; individuals who fall prey to scammers and other assaults suffer as well. Due to the increasing complexity of email spamming methods, which are moving from classic spamming (direct spamming) to more scalable, elusive, and indirect botnets for spreading email spam messages, spamming tactics are becoming more and more complicated. Spamming botnets employ a variety of sources and architectures to churn out enormous amounts of spam through email. There are thorough chronicles of spambots, which meticulously document the sequence of events and significant developments in these spambot networks' evolution. It also seeks to provide a complete review of the various email spamming botnet detection strategies that have been presented in the literature. According to both their nature and technique of detection, as well as a thorough comparison between their strengths and weaknesses, the researchers have sought to classify them. In addition, the researchers have provided an in-depth look at the effectiveness of various methods. Finally, the researchers have look at the future trends and problems in identifying email spamming botnets.

Clearly the global increase of the attacks as showcased in the work by Naser at al. [22] is demanding this research.

Henceforth, in the next section of the work, the persistent bottlenecks of the parallel research outcomes are formulated.

IV. Problem Formulation

After realizing the recent improvements over the standard methods for identification of the email types and classifying the emails into spam or ham emails, in this section of the work, the persistent research problems are identified and discussed in this section of the work.

Firstly, the during the classification of the emails, the emails are either classified as ham, that is safe email or spam, that is unsafe emails. However, the resent research outcomes demonstrate that, identification of the email sub classifications are also important for deploying the current email filters.

According to the Eq. 6, the emails are only been classified as spam or ham. Thus, this problem must be addressed.

Secondly, considering the time complexity of the process, the deployment of such email filtering frameworks is highly complex.

Assuming that the total time complexity of the existing process is T. This can be formulated in the light of Eq. 1 and 3 as,

$$T = n.n * m.m \tag{7}$$

Thus, this can be re-written as,

$$T = n^2 * m^2 \tag{8}$$

Assuming that, n ≈ m, T can be considered to be as,

$$T(n) = O(n^4) \tag{9}$$

Clearly, this is significantly high and for a large BOW[] set, the complexity can be astronomical.

Hence, this problem must also be solved.

The solutions to these two problems are proposed in the next section of this work.

V. Proposed Solution: Mathematical Models

After the detailed analysis of the parallel recent research progress and analysis of the persistent problems in the research, in this section, the proposed strategy is furnished using the mathematical models.

In the light of the Eq. 3, the BOW[] set is expected to contain multiple words, which are expected to contain all the words pertaining towards the suspected emails. It is convenient to convert the BOW[] set into a ranked sets, where the words are ranked in highest to lowest in terms of more risky to less risky. Thus, this can be formulated as:

$$R - BOW[] = \{BOW[], R[]\} \tag{1}$$

Where R[] is the associated rank set and R-BOW[] is the total set with words and the associated ranks.

Further, the threshold for each word in the BOW[] set must be calculated and the rank associated set, R-TH[x] must be formulated as:

$$R - TH[x] = \{Fq(W_X), R[X]\}$$
$$And,$$
$$R - TH[x] = \frac{\phi\{W_X \prec T[i]\{Text\}\}}{\phi\{T[i]\}} \tag{11}$$

Again, the global ranks, RG-TH[x] for each word must be calculated for the total dataset as:

$$RG-TH[x]=\{Fq(W'_X),R[X]\} = \frac{\phi\{W_X \prec LT[]\{T[i].Text\}\}}{\phi\{LT[]\}} \quad (12)$$

Finally, the emails must be classified in terms of various classes as,

$$T[i]\{class\}:R-TH[]::RG-TH[] \quad (13)$$

Henceforth, it is natural to realize that, due to inclusion of the ranks for detecting the class of the email, the proposed method does not only reduce the time complexity to $O(n^2)$, rather also provides deeper classifications of the emails.

Further, based on the proposed mathematical models, in the next section of the work, the proposed algorithm is furnished.

VI. Proposed Algorithms

After the detailed analysis of the proposed strategy in the previous section of this work, in this section the proposed algorithms are furnished.

Firstly, the BoW Collection Process using Web Crawler algorithms is furnished.

Algorithm - I: BoW Collection Process using Web Crawler (**BCP-WC**) Algorithm

Input:
U[] as set of URLs for crawling

Output:
BOW[] set as Bag or Words

Process:

Step - 1. For each element in the U[] list as U[i]
  a. Extract the list of words from the U[i] as W[j]
  b. If W[j] is noted as negative word
  c. Then, BOW[k]=W[j]
  d. Else, Continue

Step - 2. Return BOW[]

Even the most powerful search engines are only able to cover a small amount of the Internet's publicly accessible content. An investigation in 2009 found that even the most well-known search engines only index 40 to 70 percent of the searchable Web. A 1999 research by Steve Lawrence and Lee Giles found that no search engine had indexed more than 16% of the Internet at the time. Considering that a crawler gets just a small portion of the web pages, it's critical that the sites it downloads include the most relevant content.

Web pages must be prioritised based on a metric of importance. A page's significance can be gauged by a variety of factors, including the quality of its content, the number of links pointing to it, and even the URL itself (the latter is the case of vertical search engines restricted to a single top-level domain, or search engines restricted to a fixed Web site). As the whole set of Web sites cannot be known during crawling, creating an effective selection policy is made more complex.

Crawling scheduling policies were initially studied by Junghoo Cho and colleagues. They used a 180,000-page crawl from the stanford.edu domain and a variety of crawling algorithms to generate their data. To see which measures ranked highest in terms of breadth, backlink count, and partial PageRank, researchers used a variety of techniques. According to one of the findings, the best technique for crawlers looking to download pages with high Pagerank early in the crawling process is the partial Pagerank strategy. However, these findings only apply to a single field of study. Cho also completed a PhD in web crawling at Stanford University.

Secondly, the Email Text Extraction using Subsetting Method algorithm is furnished.

Algorithm - II: Email Text Extraction using Subsetting Method (**ETE-SM**) Algorithm

Input:
LT[] set as collection of emails

Output:
T[] set as collection of texts from all emails

Process:

Step - 1. For each element in the LT[] set as LT[i]
  a. Separate LT[i] into subsets using tags
  b. If LT[i].tag == body
  c. Then, T[j] = LT[i].Text

Step - 2. Return T[]

Spammer infections may have a feature that looks for email addresses on the computer's hard drive and/or network connections. These scanners unearth email addresses that have never been made public online or in Whois. Mail traffic destined for other computers on the shared network segment can be intercepted by a hacked machine on that segment. The spammer receives the harvested addresses back through the virus-created bot-net. Additionally, the addresses may be supplemented with additional data and then cross-referenced in order to obtain financial and personal information about the individuals.

As part of the "e-pending" method of direct-marketing database appending, email addresses are added. Prospect lists are typically obtained by direct marketers from a variety of sources, including subscriptions to magazines and customer lists. Direct marketers can send targeted spam email by scanning the Internet and other resources for email addresses that match the names and street addresses in their database. As with most spammer "targeting," this is not exact; consumers have claimed, for example, receiving requests to mortgage their property at a specific street location — with the address plainly being a business address containing mail stop and office number. Users have experienced similar experiences.

Thirdly, the Email Class Detection using Ranking Method algorithm is furnished.

Algorithm - III: Email Class Detection using Ranking Method (**ECD-RM**) Algorithm

Input:

BOW[] set as Bag or Words

T[][] set as collection of texts from all emails

Output:

Class of Email as {SPAM - Chain Letters, SPAM - Ads, SPAM - Spoofing, SPAM - Malware, Very positive}

Process:

Step - 1. For each text in the list BOW[] as BOW[i]

I. Build the ranking set as R-BOW[i] using Eq. 10
II. Calculate the word threshold as R-TH[i] using Eq. 11
III. Calculate the global threshold as RG-TH[i] using Eq. 12

Step - 2. For each word in the email list as T[][] as T[n][m]

I. Calculate the word ranks of T[n][m] using Eq. 13
II. If T[n][m].Rank == top 40%
III. Then, mark T[n][m] as SPAM - Chain Letters
IV. If T[n][m].Rank == top 30%
V. Then, mark T[n][m] as SPAM - Ads
VI. If T[n][m].Rank == top 20%
VII. Then, mark T[n][m] as SPAM - Spoofing
VIII. If T[n][m].Rank == top 10%
IX. Then, mark T[n][m] as SPAM - Malware
X. If T[n][m].Rank == top 90%
XI. Then, mark T[n][m] as Very positive

Step - 3. Return class of email

For the purpose of sending their messages, spammers may commit fraud. These "disposable" accounts are typically created by spammers using fake names, addresses, phone numbers and other contact details. To pay for these accounts, they frequently make use of stolen or fake credit card numbers. As the host ISPs uncover and shut down each account, they can rapidly move on to the next.

They may go to tremendous efforts in order to hide the origin of their message. Large corporations may outsource the transmission of their messages to another firm in order to deflect criticism or email blocking to a third party. Others use email faking techniques (much easier than IP address spoofing). Because the email protocol (SMTP) does not require any authentication by default, spammers can send messages that appear to come from any email address they want to use. This can be avoided by requiring the usage of SMTP-AUTH, which allows the specific account from which an email originates to be positively identified.

Due to the fact that receiving mail servers record genuine connections from the final mail server, it is impossible for a sender to spoof email delivery chains (the 'Received' header). Spammers use forged delivery headers to fool legitimate servers into believing that an email has already passed through several of their own.

For legitimate email users, spoofing can have catastrophic implications. In addition to clogging up their email inboxes with "undeliverable" emails and loads of spam, they can be incorrectly classified as a spammer. As a result, they may receive a barrage of angry emails from spam victims, as well as the threat of having their Internet service terminated for spamming.

Further, in the next section, the obtained results from the algorithms are furnished.

VII. Results and Discussions

The obtained results from the proposed algorithms are highly satisfactory and are discussed in this section of the work.

Firstly, the dataset analysis is furnished here (Table I).

TABLE I. Dataset Characteristics

Parameter	***Value***
Number of Samples	1929
Spam Emails	1543
HAM Emails	386
Minimum Length of Email Text	111
Mean Length of Email Text	231
Maximum Length of Email Text	334
Missing Words	22279
Unique Words	325287

It is natural to realize that the dataset demonstrates highly unique characteristics to be considered for testing on the proposed algorithms.

The results are also visualized graphically here (Fig. 1).

Secondly, as proposed in the mathematical models and during the algorithm design, the work summarizes the email in terms of overall scores. The details of the assumption of the scores and correlated email classes are furnished here (Table II).

Thirdly, the email classification results are showcased (Table III). The actual tests were carried out on a total of 1923 itemsets, however in this literature only 20 samples are furnished.

The obtained results are also visualized graphically here (Fig. 2).

Further, based on the ranking and matching of the email contents with various classes, the emails are finally classified in five pre-defined classes (Table IV).

Further, the results are visualized graphically here (Fig. 3).

Fourthly, the time complexity analysis of the proposed algorithms is furnished here (Table V).

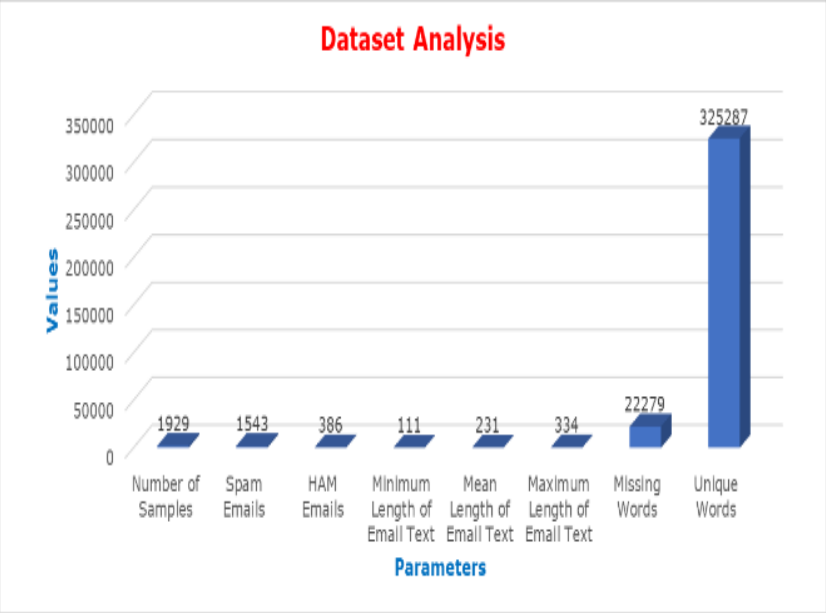

Fig. 1. Dataset Analysis.

TABLE II. EMAIL CLASS AND RANKING MAPPING

Over All Rank	***Email Class***
0	SPAM - Malware
1	SPAM - Spoofing
2	SPAM - Chain Letters
3	SPAM - Ads
4	Very positive

TABLE III. EMAIL CLASSIFICATION RESULTS

Email Seq. #	***No. of Words***	***Strong - HAM***	***HAM***	***Neutral***	***SPAM***	***Strong - SPAM***
1	37968	0%	17%	82%	1%	0%
2	1420	2%	48%	45%	5%	1%
3	4333	19%	78%	2%	0%	0%
4	2513	1%	7%	69%	21%	2%
5	2264	1%	12%	65%	20%	2%
6	278	5%	72%	21%	1%	0%
7	8207	3%	14%	43%	34%	5%
8	3361	1%	9%	79%	10%	1%
9	1617	21%	66%	12%	1%	0%
10	1184	1%	8%	78%	13%	1%
11	564	1%	17%	75%	6%	1%
12	1150	1%	6%	57%	33%	3%
13	1080	0%	7%	80%	12%	1%
14	6190	0%	1%	10%	81%	8%
15	492	8%	88%	4%	0%	0%
16	784	7%	83%	9%	0%	0%
17	907	8%	80%	8%	3%	1%
18	5153	0%	7%	90%	3%	0%
19	766	1%	6%	27%	58%	8%
20	463	4%	69%	26%	0%	0%

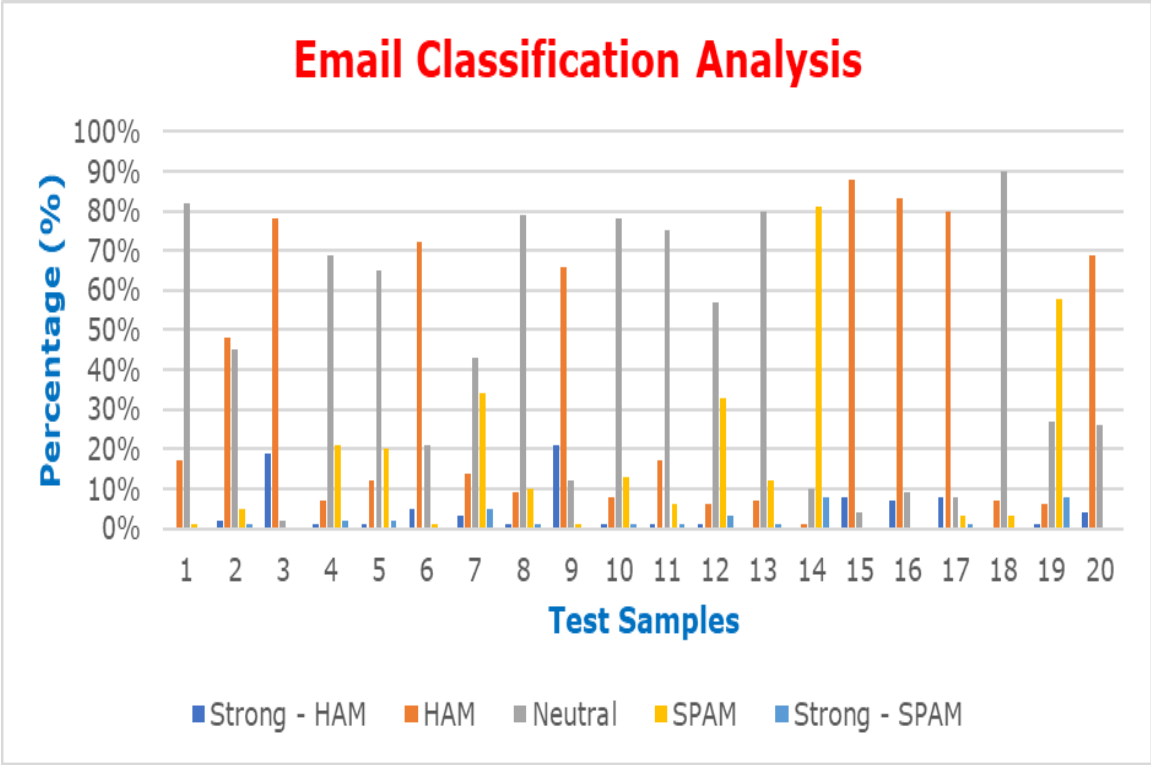

Fig. 2. Email Classification Analysis

TABLE IV. EMAIL FINAL CLASSIFICATION RESULTS

Email Seq. #	*Email Score*	*Detected Email Type*
1	2	SPAM - Chain Letters
2	3	SPAM - Ads
3	3	SPAM - Ads
4	2	SPAM - Chain Letters
5	2	SPAM - Chain Letters
6	3	SPAM - Ads
7	2	SPAM - Chain Letters
8	2	SPAM - Chain Letters
9	3	SPAM - Ads
10	2	SPAM - Chain Letters
11	2	SPAM - Chain Letters
12	2	SPAM - Chain Letters
13	2	SPAM - Chain Letters
14	1	SPAM - Spoofing
15	3	SPAM - Ads
16	3	SPAM - Ads
17	3	SPAM - Ads
18	2	SPAM - Chain Letters
19	1	SPAM - Spoofing
20	3	SPAM - Ads

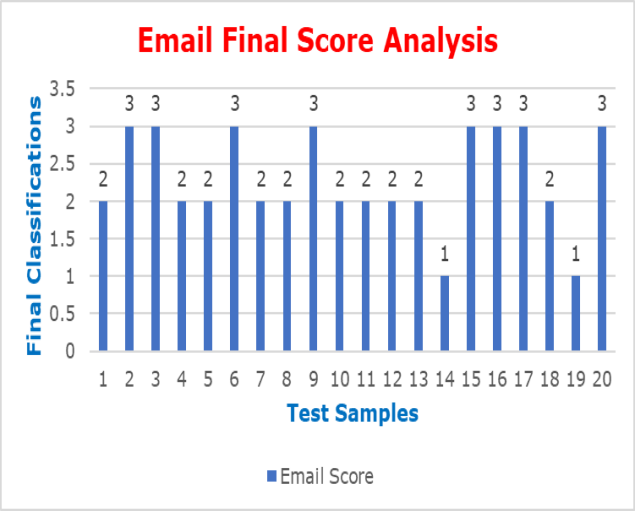

Fig. 3. Email Final Classification Analysis.

It is significant to observe that the time taken to process the complete email and further classify these emails are significantly low. The results are also visualized graphically here (Fig. 4).

Finally, the accuracy analysis results are furnished here (Table VI).

The mean accuracy of proposed system is nearly 99.42%.

Henceforth, in the next section of this work, the proposed strategy is compared with other parallel research outcomes.

TABLE V. EMAIL CLASSIFICATION TIME COMPLEXITY

Email Seq. #	*Detected Email Type*	*No. of Words*	*Time (ns)*
1	SPAM - Chain Letters	37968	164.4
2	SPAM - Ads	1420	6.1
3	SPAM - Ads	4333	18.8
4	SPAM - Chain Letters	2513	10.9
5	SPAM - Chain Letters	2264	9.8
6	SPAM - Ads	278	1.2
7	SPAM - Chain Letters	8207	35.5
8	SPAM - Chain Letters	3361	14.6
9	SPAM - Ads	1617	7.0
10	SPAM - Chain Letters	1184	5.1
11	SPAM - Chain Letters	564	2.4
12	SPAM - Chain Letters	1150	5.0
13	SPAM - Chain Letters	1080	4.7
14	SPAM - Spoofing	6190	26.8
15	SPAM - Ads	492	2.1
16	SPAM - Ads	784	3.4
17	SPAM - Ads	907	3.9
18	SPAM - Chain Letters	5153	22.3
19	SPAM - Spoofing	766	3.3
20	SPAM - Ads	463	2.0

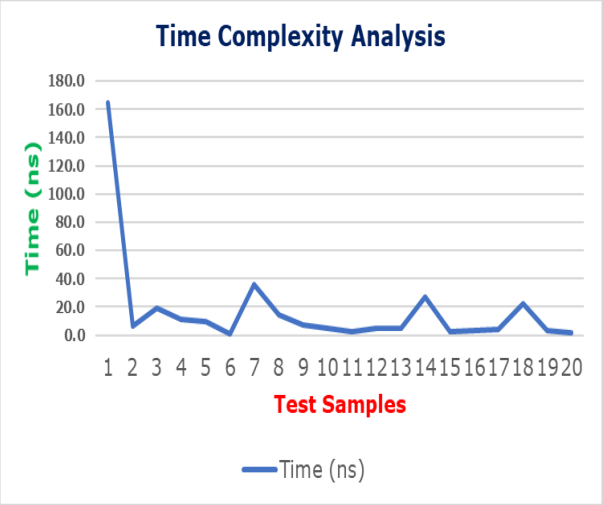

Fig. 4. Email Classification Time Complexity.

TABLE VI. ACCURACY ANALYSIS

Trial Seq #	*Number of Samples*	*Accuracy (%)*
1	412	98.943
2	425	99.224
3	468	99.637
4	488	99.997
5	136	99.344

VIII. COMPARATIVE ANALYSIS

In order to achieve the understanding, that the proposed system can outperform the other parallel research outcomes, it is important that the proposed system performances are compared with the parallel research results. Henceforth, the comparative analysis reports are furnished here (Table VII).

TABLE VII. COMPARATIVE ANALYSIS

Author & Year	*Proposed Method*	*Model Complexity*	*Accuracy (%)*
H. V. Bathala et al. [1], 2021	Filtering	$O(n^4)$	96.23
M. Hina et al. [2], 2021	Classification	$O(n^4)$	97.12
M. K. Islam et al. [3], 2021	Word Extraction	$O(n^4)$	97.68
C. Bansal et al. [4], 2021	Hybrid	$O(n^4)$	97.58
S. Sharma et al. [5], 2021	Hybrid	$O(n^4)$	98.32
Proposed Method, 2022	Classification, Ranking, Word Extraction and Deep Clstering	$O(n^2)$	99.42

Further, in the next section of the work, the research conclusion is presented.

IX. RESEARCH CONCLUSION

It's important to categorise emails because they've become one of the most common methods of communication. Because of the wide variety of ways people speak, determining whether an email is safe or not is a difficult task. There have been significant benchmarks established in email spam detection in most of the parallel studies, however. Even so, the standard spam-detection mechanism is heavily influenced by the languages used or chosen. As a result, many standard emails with legitimate text and information are incorrectly labelled as spam. In this way, the proposed machine learning method is used to solve the problem of email classification at a deeper level. Emails can only be categorised as spam or ham by the standard processes. As a result, a precise classification of the emails has thus far failed. As a result, this work proposes a novel method for classifying emails based on their severity using the proposed deep clustering process. With a 99.4 percent accuracy rate, the proposed work can detect and classify emails into a total of five categories.

REFERENCES

[1] 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), Kurnool, India, 2021, pp. 1-6.

[2] M. Hina, M. Ali, A. R. Javed, G. Srivastava, T. R. Gadekallu and Z. Jalil, "Email Classification and Forensics Analysis using Machine Learning," 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), Atlanta, GA, USA, 2021, pp. 630-635.

[3] M. K. Islam, M. A. Amin, M. R. Islam, M. N. I. Mahbub, M. I. H. Showrov and C. Kaushal, "Spam-Detection with Comparative Analysis and Spamming Words Extractions," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-9.

[4] C. Bansal and B. Sidhu, "Machine Learning based Hybrid Approach for Email Spam Detection," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-4.

[5] S. Sharma and C. Azad, "A hybrid approach for feature selection based on global and local optimization for email spam detection," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6.

[6] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "An Unsupervised Approach for Content-Based Clustering of Emails Into Spam and Ham Through Multiangular Feature Formulation," in IEEE Access, vol. 9, pp. 135186-135209, 2021.

[7] S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli and W. A. H. M. Ghanem, "Training Neural Networks by Enhance Grasshopper Optimization Algorithm for Spam Detection System," in IEEE Access, vol. 9, pp. 116768-116813, 2021.

[8] M. RAZA, N. D. Jayasinghe and M. M. A. Muslam, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," 2021 International Conference on Information Networking (ICOIN), Jeju Island, Korea (South), 2021, pp. 327-332.

[9] R. Al-Haddad, F. Sahwan, A. Aboalmakarem, G. Latif and Y. M. Alufaisan, "Email text analysis for fraud detection through machine learning techniques," 3rd Smart Cities Symposium (SCS 2020), Online Conference, 2020, pp. 613-616.

[10] S. Gibson, B. Issac, L. Zhang and S. M. Jacob, "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," in IEEE Access, vol. 8, pp. 187914-187932, 2020.

[11] I. Saha, D. Sarma, R. J. Chakma, M. N. Alam, A. Sultana and S. Hossain, "Phishing Attacks Detection using Deep Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1180-1185.

[12] N. Kumar, S. Sonowal and Nishant, "Email Spam Detection Using Machine Learning Algorithms," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 108-113.

[13] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework," in IEEE Access, vol. 8, pp. 154759-154788, 2020.

[14] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," in IEEE Access, vol. 7, pp. 168261-168295, 2019.

[15] M. Gupta, A. Bakliwal, S. Agarwal and P. Mehndiratta, "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers," 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, India, 2018, pp. 1-7.

[16] V. Vishagini and A. K. Rajan, "An Improved Spam Detection Method with Weighted Support Vector Machine," 2018 International Conference on Data Science and Engineering (ICDSE), Kochi, India, 2018, pp. 1-5.

[17] S. Chawathe, "Improving Email Security with Fuzzy Rules," 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), New York, NY, USA, 2018, pp. 1864-1869.

[18] G. Al-Rawashdeh, R. Mamat and N. Hafhizah Binti Abd Rahim, "Hybrid Water Cycle Optimization Algorithm With Simulated Annealing for Spam E-mail Detection," in IEEE Access, vol. 7, pp. 143721-143734, 2019.

[19] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues," in IEEE Access, vol. 5, pp. 9044-9064, 2017.

[20] W. Z. Khan, M. K. Khan, F. T. Bin Muhaya, M. Y. Aalsalem and H. Chao, "A Comprehensive Study of Email Spam Botnet Detection," in IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2271-2295, Fourthquarter 2015.

[21] A. N. Jaber, L. Fritsch and H. Haugerud, "Improving Phishing Detection with the Grey Wolf Optimizer," 2022 International Conference on Electronics, Information, and Communication (ICEIC), 2022, pp. 1-6, doi: 10.1109/ICEIC54506.2022.9748592.

[22] A. N. Jaber and L. Fritsch, "COVID-19 and Global Increases in Cybersecurity Attacks: Review of Possible Adverse Artificial Intelligence Attacks," 2021 25th International Computer Science and Engineering Conference (ICSEC), 2021, pp. 434-442, doi: 10.1109/ICSEC53205.2021.9684603.