International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.13, No.1 (May-23)

http://dx.doi.org/10.12785/ijcds/1301114

Advanced Spam Filtering in Electronic Mail Using Hybrid the Mini Batch K-Means Normalized Mutual Information Feature Elimination with Elephant Herding Optimization Technique

Ahmad Mtair AL-Hawamleh[1]

[1]*Department of Electronic Training, Institute of Public Administration, Riyadh, Saudi Arabia*

Received 3 Oct. 2022, Revised 30 Apr. 2023, Accepted 14 May. 2023, Published 30 May. 2023

Abstract: The danger posed by spam is quickly becoming more widespread in today's online environment. It results in a loss of money for both the companies and the users. There have been a lot of clever ideas made to filter out spam. In the first section of this study, the topic of spam is discussed, along with its features, several categories of spam, Spam strategies, various forms of spam, the drawbacks of spam, spam filtering technologies, and the effects of spam letters. When a person is enrolled with social networking sites like Twitter, Facebook, and social job networking sites, spam is more likely to have a negative impact on their online experience. There are four different sorts of spam that may be sent. Usenet spam, instant messaging spam, mobile phone spam, and email spam are the four main types. The term "USENET spam" refers to the practice in which spammers distribute advertising over a large number of newsgroups simultaneously. Spammers make use of instant messaging platforms such as AIM, Windows Live Messenger, and MySpace chat rooms to get user information and then send unwanted messages to those users. The practice of sending unsolicited text messages to those who use mobile devices is known as mobile spam. Because of spam's noisy properties and the time constraints placed on its categorization, determining the optimal spam classification algorithm has become a laborious undertaking in and of itself. The selection of features is a very important part of the classification process since using the most accurate features possible produces the highest accuracy. Optimization techniques such as modified GA, improved RBNN, s-cuckoo search, and enhanced harmony search are introduced with linear, polynomial, and quadratic kernels of SVM for spam classification. This is done in order to achieve a high level of accuracy in spam classification. The Mini batch K-Means Normalized Mutual Information Feature Extraction (KNFE) with Elephant Herding Optimization is used in the first step of the process, which is referred to as feature selection, for the purpose of selecting the pertinent features (EHO). Following the selection of features, a Radial Bias Neural Network classifier will sort the emails into those that are spam and those that are valid. When it comes to the categorization of emails, modified optimization-based feature selection produces superior outcomes than the conventional genetic algorithm. The reproduction process comes after the crossing and the mutation, which is the reason why the improvement is possible. Therefore, there is no possibility of the issue of degradation occurring since the best answer developed by the present generation will be better than those developed in the past. However, genetic algorithms employ a random method to pick parameters. Because of this, they do not perform well in situations when the population size is low, the pace of change is fast, and the fitness function must be selected with great care. It is evident from the findings that the suggested approach to spam categorization was successful in achieving better outcomes.

Keywords: Mini batch K-Means Normalized Mutual Information Feature Elimination (KNIFE), Elephant Herding Optimization (EHO), Spam Filtering, Radial Bias Neural Network.

1. INTRODUCTION

Email, which stands for electronic mail, is a time-saving method of communication that has recently gained popularity among businesses and private persons alike. E-mail has become an increasingly popular method for individuals to stay in touch with their friends, family, co-workers, clients, and business partners in today's modern world [1], [2], [3]. Spam, which is also known as unsolicited bulk email or junk mail, has become a threat that is becoming increasingly difficult to detect and is being delivered in incredibly high volumes [4]. Unfortunately, as the use of email has evolved, so have the threats associated with it. In particular, spam, which is also known as unsolicited bulk email or junk mail. Email service providers face a significant obstacle in the form of spam, which poses a possible risk to their business. For example, it is currently a difficult process to locate genuine emails inside an email inbox that is filled with spam messages. It is also a costly issue, costing service providers and organisations billions of dollars per year in missed bandwidth charges alone [5], [6], [7]. In addition to the

E-mail address: alhawamleha@ipa.edu.sa

expenses associated with bandwidth use, it is estimated that each instance of spam results in a loss of one dollar worth of productivity on the part of an organization's workforce [8]. There are a few different ways that may be taken to minimise or lessen the quantity of spam that is received by people. Legislative initiatives, such as those that prohibit spam email, are included in these strategies as one method. Other methods are referred to as Origin-Based filters, and they are based on the concept of making use of network information and IP addresses (which stands for "Internet Protocol") in order to determine whether or not a message is spam [9]. The filtering approaches, which seek to determine if a message is spam or not based on the message's content and other characteristics, are the most prevalent strategies. These techniques are used most often [10].

A. Email Communication

One of the most common ways that people communicate these days is via the use of email [11]. The sheer number of current users, which is believed to be as close to three quarters of a billion persons and is still increasing, is the finest example of the astonishingly swift adoption of this communication medium [12]. This kind of communication has the straightforward advantages of being practically quick, user-friendly, and expensive in almost no way for each individual message [13]. This protocol is the foundation of the present email system [14]. It made it possible for people to communicate with one another by exchanging messages using a system that was based on the SMTP protocol and email addresses [15], [16]. Because of these protocols, messages could be sent from one user to another, which made it possible for users of various applications to connect with one another in a way that was not reliant on the service provider or the client programme [17], [18], [19].

B. The Meaning of the Term "Spam"

Internet users are being inundated with many versions of the same message, which is known as spam, in an effort to coerce those individuals into accepting the measure even if they would not have chosen to do so otherwise [20]. The vast majority of spams are advertising for various commercial services or items, most often questionable goods, plans to get wealthy quickly, or dating services [21]. One of the most significant challenges to maintaining network security is spam [22], [23]. Sending spam has a relatively little cost associated with it for the sender, but the majority of the expenses associated with it are borne by the receiver or the service provider rather than the sender. The user wasted time while the spam email was being followed, and the ISP lost bandwidth while transporting the junk. Despite the fact that junk e-mail is the most well-known kind of spam, other forms of communication and information technology are also experiencing the same issue [24].

For the purpose of screening and preventing spam emails, DNS-Based Blacklists are used. The mail server is a potential application for this method. There are two distinct kinds of blacklists that may be used with the Domain Name System [25]. There are two types of blacklists: those based on IP addresses and those based on domain names [25]. The vast majority of DNSBLS are IP-Based, meaning that they examine the Internet Protocol address of the server that is delivering the email. When an email is sent to a mail server, the anti-spam software that is operating on that mail server inspects the IP address included in the email's header to determine whether or not the IP address is on a blacklist. If the sender's Internet Protocol address is on a blacklist, the email is classified as spam; otherwise, it is known as ham and is sent straight to the recipient. When many domains are hosted on the same IP address, domain-based blacklists are implemented [26]. Domain-based blacklists are very uncommon and are also referred to as right-hand blacklists [25]. This kind of list determines whether or not a domain name is included on a blacklist. In the event that the domain name is banned, the email will be marked as spam; otherwise, it will be regarded as ham, and it will be sent straight to the recipient.

Protecting oneself from spam-sending software that is automated requires the usage of challenge response systems [27]. This approach made the person doing the sending responsible for doing the authentication. The sender is contacted with a request by this system. In addition, the sender is asked to provide a response to the challenge. The programmes that are used to transmit spam automatically are unable to respond to this challenge [28]. When the spam came from a mailbox that was functioning properly or was being watched by the spammer, challenge-response filters would be more effective in identifying and filtering spam. This filter is applied to the server that is sending the data as well as the server that is receiving the data. A method known as nation-based filtering is used to prevent email from being sent from particular countries [29]. This decision is made based on the IP address of the sender, which reveals the country from which the email originated [29].

The process of grey listing is a method for temporarily rejecting communications from mail servers that are unknown as the sender [30]. Spam bots are specialised pieces of software that may send hundreds of unwanted emails in a very short period of time. The system will generate a tupla sender-receiver pair if it gets an email from an unknown sender that is not included in a white listing. When the tupla appears for the first time in the system, the email is flagged as spam and is thus sent directly back to the sender. That email will be sent again if the server is legitimate [31] As a result, the algorithm locates the tupla for the second time, and the email is then deemed trustworthy and sent to the recipient. These spam bots operate differently from conventional email servers, and they do not adhere to the RFC requirements for email. Based on the content of messages included inside emails are analysed by filters in order to identify spam. In this part of the paper, the content-based approaches, such as heuristics filters and machine learning filters, were explored. Rule-based filters are referred to as heuristic filters. These filters look for patterns

that are known to be present in spam emails, such as certain words and phrases, exclamation marks, malformed message headers, and capital letters. These patterns are then used to categorise spam emails. Using a predetermined set of criteria that were hand-coded, it examines the contents of a message and determines whether or not it is spam. Techniques for filtering content depend on the provision of lists containing words or regular expressions that are not permitted in e-mail communications. It also does an analysis on the email header, which contains information such as the list of recipients, the originating IP address, and the topic [32].

In the checksum-based filtering approach, filters remove anything that may potentially differ between messages, reduce what is left to a checksum, and then search that checksum up in a database that gathers the checksums of messages that email consumers perceive to be spam [33]. The spammers have the ability to inject one-of-a-kind hash busters that are invisible into the body of each of their mails [33]. As a result, each message is distinct and has its own unique checksum. In order to train a spam filter, a machine learning technique is necessary, and this approach needs a huge dataset consisting of both spam and ham emails. The machine learning strategy does not need any explicit rules to be specified in order to be implemented. Instead, what is required is a collection of papers that have already been classified—these will serve as training examples. There are many different types of spam filters to choose from. These methods include clustering strategies, decision trees, statistical filters, genetic algorithms, artificial immune systems, and artificial neural networks [34], [35]. The statistical filtering method divides incoming emails into a number of tokens on their own and then uses these tokens to search for matches in a database [36], [37]. It determines if an email is junk mail or ham [36], [37] Bayesian filtering, which relies on word probabilities, is one of the most well-known approaches to preventing spam. This approach, which is a component of statistical methodology, breaks the income message up into terms that are referred to as tokens and analyses the frequency with which each token appears in spam emails and ham emails. Spammers use a strategy called Bayesian poisoning to undermine the effectiveness of this approach by injecting lines of meaningless random words. A technique for determining whether or not something is spam is called the Chi by degrees of freedom test [38]. In the process of determining whether or not an email is spam, a text classification approach called Support Vector Machine is used [39]. It is a way of learning via supervision. The AdaBoosting technique is a method that boosts the weaker learning algorithm that is used to efficiently filter spam [40], [41]. When improved with a boosting algorithm, Bayesian and decision tree approaches performed well in the task of spam filtering [40], [41]. The Maximum Entropy Model is an additional approach to machine learning. It is a tried-and-true model from the field of natural language processing that is put to use in the elimination of spam.

The Artificial Neural Network is a well-known model that was developed specifically for the purpose of identifying spam [42]. It will categorise incoming emails according to characteristics that are typical of emails. Some examples of artificial neural networks (ANNs) include perceptrons, multi-layer networks, and learning vector quantizers. In order to locate a linear function of the feature vector, perceptrons are a useful tool. A multi-layer neural network is a network that is composed of many layers of linked perceptrons that are arranged in a hierarchical fashion [42]. An example of a nonlinear classifier is the multilayer perceptron. Text categorization is an application that lends itself particularly well to the use of learning vector quantizers.

An artificial immune system is a technique of machine learning that is used in the battle against spam and computer viruses [43]. The immune systems of living species provide the foundation for this approach. This approach is used to categorise spam or ham based on artificial lymphocytes made from a gene database [43]. The genes represent miniature languages that contain keywords that are checked in spam, and the technique is based on artificial lymphocytes created from a gene database [44]. KNN Filtering An instance-based classifier is known as the K-Nearest Neighbor algorithm. The number of neighbours who contributed to the categorization is indicated by the k parameter [45]. The email is first transformed into a vector with a large number of dimensions, and then the approach calculates the distance between the vectors of each email. A text's k-nearest neighbour in feature space may be determined using this approach [45]. Vectors that are physically near to one another and their neighbours combine to create clusters. When it comes to filtering email, this approach is straightforward and divides emails into two categories: genuine email and spam.

The support vector machine is a kind of algorithm that is used for classification and regression [46]. IRBNN is used so that the ideal hyper plane may be found. The notion of decision planes, which establish decision boundaries, serves as the foundation for this approach. This classification splits the data into two distinct classes that are then formed from the training examples. This helps to ensure that the gap between the two classes is as wide as possible. The classification of spam messages, in particular, benefits from the use of this strategy, which is successful in categorising text in general.

A decision tree may be used to classify data [47] The process of data mining makes extensive use of this methodology [47]. It screens communications sent through email. This tree has a limited number of branches. The process of classifying a document begins at the root node and works its way outward, picking conditions from branch nodes based on whether or not they were met. There is a choice at C4.5. A tree may be used as a classifier. A binary tree is produced as a result. The observations are

represented by the nodes in the interior of the tree, while the conclusions are represented by the nodes on the tree's leaves. Methods of collecting email addresses that are used by spammers get the email addresses of internet users from a variety of different sources and create a significant amount of frustration for those people. The email address was stolen by spammers using all three of the following methods: They are gathering the email addresses of the users by using spam bots, carrying out dictionary attacks, and acquiring address databases from people or organisations. Web spiders were used by spammers in order to harvest email addresses from various websites. Spammers have also been known to get email addresses straight from the results of Google searches. Spammers send out their unwanted messages by using a method known as automated bulk mailers, which is a computer software. This software requires very little time investment yet is capable of sending a significant number of emails. Spammers may also transmit their unwanted messages by using zombie networks. Spammers have the ability to conceal their own email address. A way of obfuscating the message content was employed by spammers. This strategy left the term accessible to humans but made it less likely that a literal computer programme would identify it. Emails sent in HTML, on the other hand, provide the spammer with more tools to disguise the content.

Spammers not only steal sensitive financial information from their victims, but they also slow down their computers by introducing viruses. The downloading of spam emails eats up the available bandwidth on the network, wastes memory and the user's time, reduces productivity, and resulting in monetary loss for both the user and the enterprise. In addition to this, it has an impact on the general performance of using the internet. Every person who uses the internet should take a few seconds out of their day to read and remove any spam emails that are in their mailbox. The purpose of spammers is to get a response from a small percentage of internet users so that they may maximise their financial gain by sending mass emails to those who use the internet. The photos that are sent by those who spam contain something called a web bug that is placed inside them. It monitors and compiles information about the time, location, and IP address of the computer used by each receiver in order to determine when and where emails are viewed. The user will have a tough time seeing the item that has been inserted into the picture. Individuals have access to a variety of methods that may limit the accessibility of their email addresses, so lowering or eliminating the likelihood that spam will be sent to those addresses.

- One strategy for reducing the amount of spam received in one's inbox is to restrict the distribution of one's email address to a select few contacts.
- When you are forwarding messages, it is a recommended practice to place all of the recipients' names after the "bcc:" field rather than after the "to:" field.
- Because of this technique, the situation in which email addresses are used for the purpose of spamming is avoided, and the chance that the address will be transmitted by machines that have been infected with malware that harvests email addresses is also reduced.
- One technique to prevent having one's email address harvested is to post in secret or with a false name and address. However, users should check to be sure that the phoney address they use is not a real one.
- Anyone who uses the internet should refrain from replying to spam mail.
- Users of the Internet should refrain from filling out contact forms. Users of the internet should avoid downloading documents in the HTML format since it takes the user's personal information. It is possible for HTML-written spam to include web bugs, which provide the sender with the information that the recipient's email address is correct and that the message has not been blocked by a spam filter.

In order to filter spam, there are many different ways that have been deployed. However, spammers make use of obfuscation and a variety of other strategies in an effort to circumvent anti-spam filters. In contrast to the earlier work done by the researchers, this approach is unable to differentiate between spam and ham. However, this study differentiates spam from other types of spam. In the study that was suggested, an Improved Radial Bias Neural Network was used to categorise spam. The work that was done classified spam emails using a technique called an Improved Radial Bias Neural Network. This piece of work comprises a database that contains a list of spam terms as well as the email addresses of spammers. This approach pulls characteristics from the email, which are then compared to a list of spam features that have been saved in the database and rated according to their values. The results of this comparison are then used to classify the words and addresses in accordance with the rating. The improved Radial Bias Neural Network system has successfully identified the junk emails as either the least risky, moderately harmful, or the most dangerous spam mail.

2. BACKGROUND OF THE STUDY

Provided a comparative analysis of several spam filters and spoke about the many machine learning methods that may be used to filter spam [48]. Their research focuses on the study of various automated filtering and machine learning approaches, such as algorithms that are rule based, content based, customised, collaborative, support vector machine, and kernel based. They offered a comparative study on several filtering methods as well as the benefits associated with each one. [49] created a system for screening spam email from Internet service providers in spite of

the tremendous traffic on the internet. They tested their approach using data collected from one of the most significant commercial internet service providers in China and found that it was successful in analysing email traffic statistics. They were successful in reducing the volume of junk mail traffic by 70.4% as a consequence. A parameter is set for the email category using the fingerprint approach, which is used to identify emails that were sent previously that were comparable. The database for mail and the database for fingerprints are both used to store information. By just adding the record in the MD, and deleting any messages that are deemed to be unnecessary. They gave an explanation regarding the three benefits that BMTC offers. They have a high level of accuracy in recognising emails, automated hand-free deployment, an online update mechanism, and the ability to handle a huge number of data while using a minimal amount of memory and an acceptable amount of CPU time.

[50] suggested doing a comparison research for the purpose of email categorization. In order to filter spam from datasets of emails, many classifiers, including Neural Network, SVM, Naive Bayesian, and J48, are used. Data preparation, data training, and data testing are the three components that make up a neural network. Extracting more informative features while deleting characteristics that are unnecessary or superfluous is what feature selection does. The pre-processed characteristics are inputted into the NN, and the NN is responsible for the generation of the email classifier. The effectiveness of NN is evaluated by the usage of the email classifier in the third phase of the testing process. In order to carry out the experiment, an error back propagation method was used. Learning and generalisation are two areas in which SVMs perform very well. SVMs learn from examples, and each example is made up of a certain number of data points followed by a label that is organised into one of two classes. They are represented by the numbers +1, which stand for one state, and -1, which stand for another state. The two classes may be distinguished using the optimal hyperplane. The Support Vector algorithm reduces the gap between the nearest +1 and -1 point as much as possible. It creates two distinct classes that are based on the training examples provided. The support vector machine (SVM) does away with the need for a separate hyper plane by devising a method that maximises the difference in score that can be achieved between two classes. The Bayesian theorem and the notion of total probability provide the foundation for an efficient classifier known as the naive Bayesian classifier. The J48 decision tree is used to classify legal messages from spam messages using a binary tree that it generates. They assessed the performance of the four classifiers using a variety of datasets and characteristics.

$$Accuracy = \frac{Ture\ Positive\ Classified\ mail}{Total\ Number\ of\ mail} \times 100 \quad (1)$$

Precision and recall were the performance criteria for email categorization that were used for measuring how well the system worked. They hypothesised that J48 and NB classifiers achieved more accurate results than SVM and NN classifiers did.

[51] put out the idea of conducting a survey on learning-based approaches to spam filtering. In this paper, several learning-based techniques for spam filtering, such as keyword filtering, image-based filtering, language-based filtering, filters that are based on non-content aspects, collaborative filtering, and hybrid approaches, were reviewed. They evaluated and compared the findings that were produced via the many different types of filtering processes and reported their findings.

[52] suggested a system for screening spam email that has a high accuracy rate while also having a low time complexity. In order to further their investigation, they obtained Turkish mails. They used the PC-KIMMO system, which is a morphological analyser, to extract the root forms of words as the input and create the parse of words as the output. A heuristics and the n-gram technique are the foundations of this methodology. They came up with two models: one for classes in general and one that was tailored to emails specifically. The bayes rule is used by the general model to determine whether or not the email is spam or authentic. The right class of a message is identified by the second model, which does so by comparing the current message to an earlier one that is quite similar to it. The third model is one that combines and refines different perceptions. It is an amalgamation of the two types described above. For the purposes of the n-gram model, the arranging of the words in fixed order is done using free word order. This technique for preventing spam is based on categorising the text contents and raw contents of emails, with the goal of producing results from the classification of data sets. When dealing with a greater quantity of words, they encountered a difficulty that was more difficult and time-consuming to solve. The AdaBoost ensemble method is used in order to evaluate and contrast its prior work. They carried out extensive experiments on several number datasets of varying sizes and starting word combinations. They have accomplished this by achieving a high percentage of success in both the Turkish and the English languages. A multiobjective evolutionary algorithm for the filtering of spam was developed by [53].

They considered the ideas of dominance and pareto optimality. For the purpose of screening spam emails, SPAM-NSGA-II-GP is used. MOEA is a tool that is used to learn a collection of queries with high recall and accuracy. For the purpose of spam filtering, PUI datasets are used. SPAM-NSGA-II-GP, which has highly stringent filtering criteria, is used to block all of the valid emails, which are then tagged as spam. This method has a high recall but a poor accuracy. They labelled the smallest possible percentage of spam emails using the weak filtering criteria, which had a

high degree of accuracy but a poor recall rate.

The method of steganography consists of a number of significant components, the most important of which are feature extraction and categorization based on feature sets. Because of the large dimension of the feature sets utilised for steganography, the classification procedure is both difficult and time intensive. A unique feature selection technique known as FS-SDS was presented for use in steganography by [54]. FS-SDS is a wrapper-type feature selection technique that uses stochastic diffusion search to pick a reduced feature set (SDS). The general population-based search approach known as the stochastic diffusion search has been effectively implemented in this study for the purpose of steganalytic feature selection. Experiments are carried out on two distinct feature sets in order to demonstrate the utility and efficacy of FS-SDS. The similarity computation uses the suggested vector space model that is part of the document clustering technique. The vector space model is modified, and then a meta-heuristic method is layered on top of it. This is done in order to improve the usability. In order to determine the level of similarity between the various documents, the document clustering technique that uses a modified vector space model looks for the most important component of each document's vector space. Using an artificial neural network, [55] offer a technique for the categorization of text-and image-based spam emails (ANN). Training and testing will be carried out on two different data sets; the first will include text-based emails, while the second will contain image-based emails. The use of an OCR tool is necessary in order to extract text from images. Filter spam emails on the iPhone using a unique algorithm called the artificial bee-based decision tree (ABBDT) method [56]. Decision trees are used in the ABBDT technique that has been presented for the purpose of filtering spam emails for the iPhone. In addition to this, the artificial bee method is implemented in order to improve the testing precision of the decision tree. There are around 504 emails, and each of them is sorted into one of 12 categories. In order to test how well the suggested ABBDT technique works, another spam base dataset is used. This dataset was collected from the repository of machine learning datasets at the University of California, Irvine.

During the course of the literature review, it became clear that there are a great number of strategies for combating the issue of spam; all of these strategies are now being implemented in a great number of distinct types of spam filters. The majority of filters are used during the analysis of email envelopes, while a mix of heuristic and Bayesian approaches is utilised during email content analysis. In their most basic form, all of these filters sort incoming emails into two categories: spam and non-spam. The majority of spam filters determine whether or not an incoming email is spam by looking at certain terms in the data section or subject part of the email [19], [57]. as well as the source from which the email is coming. They then file the email under the appropriate category. However, this does not provide sufficient evidence to classify email as spam. Despite the availability of a great number of methods for identifying spam, user inboxes continue to be overwhelmed with unsolicited messages. The categorization of emails is difficult due to the enormous and diverse number of variables included inside the dataset, as well as the sheer volume of emails themselves. When there are a greater number of traits, most messages cannot be differentiated from one another. In many email datasets, only a small fraction of the total features are helpful in categorising emails, and utilising all of the characteristics may have a negative effect on performance. Using all of the features may also cause errors. The following goal has been taken into consideration in the development of this work:

In order to detect spam emails by utilising spam terms and the spammer's address as ranking criteria, an improved radial bias neural network classifier will be used

3. PROPOSED WORK

An email has two parts: the header and the content, sometimes known as the body. The fields that make up the header section are titled "From," "To," "CC" and "BCC," which stand for "carbon copy" and "black carbon copy," respectively, and "subjects." In the genetic algorithm with modifications, the header is not significant at all, and only the body component is considered. Words are taken from the actual transmission of the email. During the process of extracting the words, articles such as "a," "an," "the," and "for" as well as numerical values are omitted. In the evolutionary algorithm that has been updated, the first step is to establish a database that will sort ham and spam emails, and this database may be segmented into several groups depending on our preferences. It is essential to keep in mind that the number of words included in the data dictionary will grow proportionately with the size of the database. The manner in which the emails are categorised will determine which categories are selected. However, if there are fewer categories established, it is still possible to determine whether or not an email is junk mail; the risk of getting a false positive or negative result, however, rises. Chromosomes are built for the incoming messages after they have been created. The process of using a modified genetic algorithm to choose candidates for a crossover begins. As was just said, there are many different approaches one may use while doing a crossover. In crossover, the exchange of genes can only take place between members of the same group. In our approach, the multi-point and single-point crossover operations are carried out concurrently, and the bit locations are determined in a random manner. Only 12 percent of chromosomes are transferred across from one generation to the next. The next phase is mutation, which is done here to recover some of the lost genes or in our instance to recover some of the lost data. However, just three percent of the genes in this population have been altered. After that, another calculation of the fitness function is performed. Then, in order to locate the gene that matches,

the weight of the words of gene in the testing mail and the weight of the words of gene in the spam mail prototype are compared. If the number of matching genes is more than or equal to three, the spam mail prototype will be awarded one point toward its total score. If the score point is higher than a threshold score, for example 62 points, then the email is deemed to be spam mail. However, the moment at which the threshold is applied may be manually adjusted in order to get the desired effects. The fitness function that is being employed here is derived from the outcomes of experiments, and this fact has to be kept in mind.

A. *Mini batch K-Means Extraction of Normalized Mutual Information Features*

The search phrase is compared with the clusters rather than the document sets, which is one of the many ways that clustering helps to improve the overall quality and effectiveness of search engines. Because of this, the results of the search may be readily organised in accordance with the relevancy of the results. When a user searches for information on the internet, retrieval results may be delivered quickly because text document clustering groups a collection of documents according to the degree to which the content of the documents is similar.

Experiential evidences have indicated that applications for information retrieval may benefit from document clustering, and it has been utilized as a method to enhance the performance of information retrieval. Clustering has also been used to increase the performance of document retrieval. The process of document clustering is now the subject of investigation by a number of scholars, although it must yet be examined. The establishment of a collection of groups that faithfully exemplifies the subjects covered by a given document set is one of the primary objectives of the document clustering process. The success of the clustering process is dependent on selecting the appropriate technique for clustering as well as the similarity measure. The cosine similarity functions or the Euclidean distance are the measures of similarity that are used. Both of these distance measures consider the frequency with which the phrases appear in the document collection. Because every document has its own unique words, the document term matrix that was developed ended up being highly dimensional and somewhat sparse. Because of issues with large dimensionality and the sparse nature of this matrix, the performance of clustering algorithms will significantly deteriorate in the near future. The accuracy of information retrieval, measured in terms of precision and recall, is negatively impacted when documents include irrelevant attributes. Additionally, it becomes more difficult for a clustering algorithm to properly group together documents that are similar.

Dimensionality reduction, often known as DR, is a technique that may be used on this document's term matrix to decrease the dimensionality of the data. This technique is widely regarded as the most significant step in simplifying the clustering process. Through the use of DR, the big, sparse, high-dimensional matrix may be reduced in size to one that is more manageable. DR may be accomplished by either the extraction of features or the selection of features. The process of feature extraction reduces the number of dimensions of high-dimensional data, which results in the creation of a new set of features derived from the features of the original set. Dimensionality reduction is accomplished by selecting a subset of the original representation's features using quality criteria such as information gain or chi-square. This process is known as feature selection. Mini batch Clustering of text documents may be accomplished effectively with the use of a method called K-Means Normalized Mutual Information. This research discusses the usage of K-Means with feature selection in the clustering of a dataset consisting of text documents and demonstrates how it enhances the performance in terms of accuracy when compared to K-Means that does not employ feature selection. The conventional vector space model is used by both of these different clustering algorithms. The words and documents that make up the vector space model are laid up in the form of a matrix. It also includes a local and global weighting mechanism.

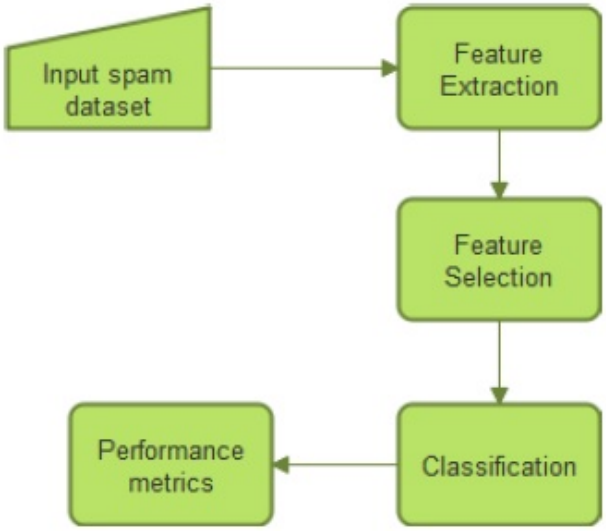

Figure 1. Basic Structure of Processing Unit

MacQueen invented the Mini batch K-Means Normalized Mutual Information method in order to find a solution to the well-known clustering issue. This technique is currently one of the most popular and efficient unsupervised learning algorithms available. The K-Means algorithm's goal is to cluster a group of objects according to the attributes that distinguish them from one another, with the number of clusters determined by the variable k. The importance of the centroid to the K-Means method cannot be overstated. The mean value of the items in the cluster is represented by the centroid value. The primary objective here is to identify k centroids, one to represent

each cluster. The formation of a cluster's centroid takes place in such a manner that it has a similarity function that places it in close proximity to other elements of the cluster. The Euclidean distance between each item in a cluster is what is used to determine how similar two things are to one another. Both the DR approach and the Ontology-based clustering method that this study paper employs make use of an algorithm called the Mini batch K-Means Normalized Mutual Information. The dataset including spam was used for all three of the aforementioned methods. The document term matrix, the number of clusters, the dataset, and k are the components that are taken into consideration by the K-Means method. This K-Means algorithm produces a set of clusters as its output, along with the number of documents contained within each cluster, the number of iterations performed, the amount of time necessary to construct the model, the sum of squared errors, the confusion matrix, and the number of documents that were improperly clustered. The K-Means algorithm's stages are laid out in Algorithm 1, which can be found here.

Algorithm 1 K-Means Normalized Mutual Information Algorithm

K-Means clustering is a statistical technique
Dataset, K as the number of clusters, Document word matrix, and similarity measure are the inputs. Measure
Clusters C1, C2, etc., and a confusion matrix are the products of this analysis.
Perform an initialization of the parameters.
Define K.
start by picking the initial set of centroids at random, and then you will apply the Mini batch K-Means Normalized Mutual Information method to the dataset.
In order to establish the connection between the clusters and the classes:
Calculate the Euclidean distance between each cluster's centre and each class's centre.
Choose the assessment entitled "Classes to Clusters."
Think about all the many jobs that are available.
Compute total distance for all situations.
Obtain the most desirable task.
Return to the third phase, and end when there are no more new assignments

Techniques for selecting features in most cases, the pursuit of optimum subsets is geared toward one of these two goals: a) decrease the number of characteristics that are picked while yet achieving some minimum degree of capacity for categorization, or 49 b) Maximize the capacity to classify for a defined subset of cardinality. The procedure for selecting features may be made more effective by improving its subset selection approaches with the help of various well-known optimizers. This will result in the procedure taking up less time. The Genetic Algorithm is a subset of the evolution-based optimization problems methods that concentrate on applying selection, mutation, and recombination to a population of competing problem solutions. This subset is part of the evolution-based optimization problems techniques. The use of GAs, which are optimizers that work in a parallel, iterative fashion, has shown to be effective across a wide range of optimization issues, including a great deal of pattern recognition and classification work. EHOs have also been used to determine an ideal collection of feature weights that enhance classification accuracy. This technique has been shown to be a successful computational approach, particularly in circumstances in which the search space is uncharacterized, not well understood, or extremely dimensional. Therefore, in order to improve the process of selecting features for the IRBNN classifier, we use EHO as an optimizer.

The social behaviour of elephants that live in herds served as an inspiration for the EHO. As is the case in any other kind of social culture, the majority of the domestic duties in a family are performed by the family's female members. In a same fashion, the elephants that inhabit EHO are organised into clans, each of which is headed by a female elephant that is often referred to as a "matriarch." The remaining members are made up mostly of junior females and young calves. When male calves reach maturity, it is natural for them to separate from their clan and have independent lives, therefore they are no longer considered members of that clan. In light of the above, the social structure of elephant herds may be summed up as follows in order to simulate their behaviour.

- Elephants are social animals and live in family groups headed by a matriarch. It is assumed, for the sake of simplicity, that each clan is composed of an identical number of elephants at all times. The status of individual elephants is revised based on their relationship to the matriarch.
- Eventually, an adult male elephant will separate himself from the herd. For the sake of modelling, the matriarch is believed to be the healthiest individual, which determines the position of the other elephants, and a separated male elephant is regarded to be the worst alternative for the group. As of this point forward, the solution j in each clan ci is updated by its current location and matriarch ci via the updating operator, and the population diversity is increased in a later generation by the separating operator.

The method may be expressed mathematically in the following manner:

The starting population is a representation of the whole number of likely solutions, and it is then partitioned into n separate clans. Because the matriarch ci has the highest level of fitness, each solution j in the ci clan shifts its place in accordance with the equation given below in 2.

$$x_{new,d,j} = x_{ci,j} + \alpha\left(x_{bst,di} - x_{ci,j}\right) \times r \qquad (2)$$

The variable denoted by is a scale factor with a value range of 0 to 1. This points to the impact that ci, the matriarch, had on $x.ci$, $j.r = 0, 1$ is a random variable that follows a uniform distribution, on the other hand.

The previous position of j in clan c j is denoted by the notation $x(mav, i, j)$. $X(los, di)$, on the other hand, is the most optimal solution for clan c 1. In addition, the phrase that is shown below is what is employed to bring each clan's fittest solution up to date.

$$x_{new,ci,j} = \beta \times x_{centre,ci} \tag{3}$$

Here, "0,1" represents the impact factor that x "convre," has on the revised solution. The location of the centre of clan ci may be determined using:

$$x_{centre,ci,d} = \frac{1}{n_{ci}} \times \sum_{j=1}^{D} x_{ci,j,d} \tag{4}$$

D is considered as the total dimension of the search space and $1 \leq d \leq D$ represents d^h. dimension. For separation of worst individual the representation is:

$$x_{warst,i} = x_{\min} + (x_{\max} - x_{\min} + 1) \times rand \tag{5}$$

The fitness level of x is the lowest in clan ci. On the other hand, "$rand$" refers to an arbitrary integer between 0 and 1 that is selected using a uniform distribution. The limits of the particular position are denoted by the values of max x and min x respectively. The EHO algorithm has its own set of benefits, one of which is that it requires a smaller number of control parameters. However, the likelihood that the solution that was randomly picked is a good solution is exactly the same as the probability that it is a terrible solution. As a result, the new candidate solution does not necessarily hold the promise of being a solution that is superior to the one that came before it. Because of the presence of these random values, the search operator does not consider the knowledge on the optimal solution or any other solutions that would have a beneficial impact on directing the EHO toward more promising regions of the search space.

This is especially the case when the separation operator is being generated in Equation 4. In addition, the value of in Equation 5 is always the same and ranges between 0 and 1. It does not change from one generation to the next. However, if it is turned into an adaptable parameter throughout the process of evolution, the solution may have a better chance of success in EHO. Within the context of the original EHO algorithm, it is seen as a constant number that ranges between 0 and 1. However, the value may be adjusted over a number of generations based on a linear function as indicated below, and this change can be included into the algorithm code.

$$\alpha = \alpha + \frac{\alpha_{\max} + \alpha_{\min}}{n} \tag{6}$$

Where _min and _max are the minimum and maximum allowable values for the range of a, respectively.

At each iteration, the value of the subject to continual adjustment, which results in an improvement in the efficiency of the initial EHO.

Initialization
Limits on the number of generations, the total population, and the borders
Generation of the population by random means
Determine the elephant's capacity for the solution.
Repeat
Classify each of the elephants according to their level of health.
Clan update;
for do ci = 1 to $nclan$ (for all clans in elephant population)
for do j = 1 to nci (for all elephants in clan c)
if then xci, j is equal to $xbestci, i$, then you should update xci, j (the previous version) and create xn, ci, j (new)
Equation 2
else
Update xci, j (old) and create xn, ci, j (new)
Equation 1
Exit conditional
Exit for j Exit for ci
end if
end for
end for
separating; with regard to ci = 1 to $nclan$ (all clans in elephant population)
Find a better elephant to take its position in Clan C. The end of eq. 4 for ci
Conduct population analysis using the new positions.
until an infinite number of generations have passed.

B. *Methods of Improved Radial Bias Neural Network Classifier*

$$V_{LFM}(t) = \begin{cases} a \ \exp(j\pi t^2); & 0 \leq t \leq T_0 \\ 0; & T_0 < t < T_{PRI} \end{cases} \tag{7}$$

$$v_j(t) = v(t - nT_{PRI} \ for \ 0 \leq t \leq T_0 \tag{8}$$

$$v(t) = \{v_j(t)\} \exp(j2\pi f_c t) \tag{9}$$

$$\tau_m = \tau_0 - \frac{2}{C}\{vmT_{PRI}\} \tag{10}$$

$$p_n(t) = \{v_j(t-\tau_n)\}\exp(j2\pi f_c(t-\tau_n)) + K_m(t) \quad (11)$$

Where, $k_m(t)$- additive thermal noises Returning signals $p_n(t)$ f basebands when depicted mathematically, form Equation 6

$$p_m(t) = \{v_j(t-\tau_n)\}\exp(-j2\pi f_c\tau_n) + k_m(t) \quad (12)$$

Which implies $P_n(f)$ can be written as Equation 12

$$P_n(f) = |V_{LFM}(f)|^2 \exp\exp(-j2\pi f_c\tau_n)\exp\exp(-j2\pi f\tau_n)+k_m(f) \quad (13)$$

Where, $V_{LFM}(f)$ represents Fourier transforms LFMSs Sampling frequency $l = [0, 1, \cdots, L-1]$ with interval Δf, and dividing by $|V_{LFM}(l\Delta f)|^2$ yields the following Equation 13.

$$P_n(f) = |v_{LFM}(f)|^2 \exp(-j2\pi f_c\tau_n)\exp(-j2\pi f\tau_n) + k_m(f) \quad (14)$$

Where $k(n,l)$ represents thermal noise's discrete samples. Substituting τ_n from Equation 4 results in Equation 9,

$p(n,l) = \exp(j2\pi n f_d T_{PRI})\exp(-2\pi l\Delta\tau_0)\exp\left(j2\pi f_d ml\left(\frac{T_{PRI}\Delta f}{f_c}\right)\right)+ k(n,l)$

$$x = x_{\min} + (x_{\max} - x_{min} + 1)rand \quad (15)$$

The accuracy is achieved better by inertia weight which is decided by evolution speed of each particle and aggregation degree of the swarm. Large inertia weight enhances global search while small inertia weight results in faster convergence.

4. EXPERIMENTAL RESULTS

This section presents the experimental results obtained from various classifiers such as KNN, Naive Bayes, and SVM with proposed methodology.

A. Classification Accuracy

Table 1 shows the classification accuracy obtained from various classifiers

TABLE I. Classification Accuracy (%) using Improved RBNN

Data Sets	KNN	Naive Bayes	Improved RBNN		
			Linear Kernel	Polynomial Kernel	Quadratic Kernel
Ling spam	79.35	85.21	95.74	93.87	93.78
Enron	78.22	84.98	94.65	93.06	92.95
Spam Assassin	77.63	86.54	94.79	92.66	92.19
CSDM C2010	78.51	81.65	92.97	91.98	91.90
PU Dataset	75.24	79.35	96.54	93.65	95.48

The following Figure 2 shows the graphical representation of the proposed method for classification accuracy, which proves that the proposed method achieved better accuracy using SVM classifier. The accuracy is achieved better by inertia weight which is decided by evolution speed of each particle and aggregation degree of the swarm. Large inertia weight enhances global search while small inertia weight results in faster convergence. In each iteration, best particle is being chosen according to their fitness values and the number of particle selected is also updated if it has a reduced value. The velocity of the particle in each dimension increases rapidly and tends to select higher number of solutions for obtaining high classification accuracy. In Ling Spam dataset, the improved RBNN method achieves 93.78% of accuracy in quadratic kernel SVM while simple RBNN achieves only 92.11% and 92.94% of improvement is achieved. In Enron Spam dataset, the improved RBNN method achieves 92.95% of accuracy in quadratic kernel SVM while RBNN achieves only 85.07% and 89.01% of improvement is achieved. In Spam Assassin dataset, the improved RBNN method achieves 92.19% of accuracy in quadratic kernel SVM while RBNN achieves only 83.23% and 88.01% of improvement is achieved. In CSDMC2010 dataset, the improved RBNN method achieves 91.90% of accuracy in quadratic kernel SVM while RBNN achieves only 85.57% and 88.73% of improvement is achieved. In PU dataset, the improved RBNN method achieves 95.48% of accuracy in quadratic kernel SVM while RBNN achieves only 91.52% and 94.03% of improvement is achieved.

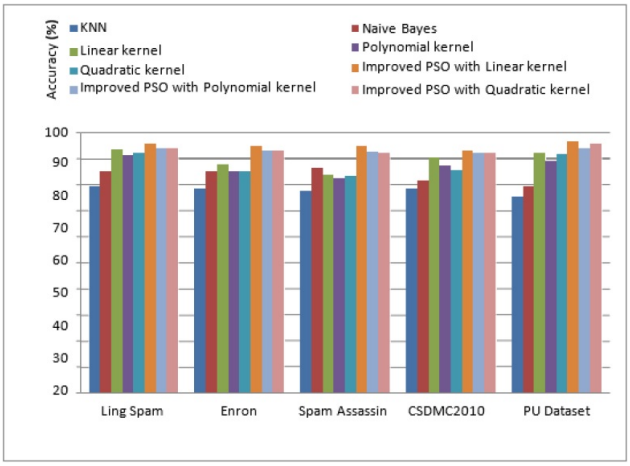

Figure 2. Classification Accuracy using Improved RBNN

B. Execution Time (seconds)

Table 2 shows the execution time of different classifiers such as KNN, Naive Bayes, and SVM.

TABLE II. Execution Time (seconds) using Improved RBNN

Data Sets	KNN	Naive Bayes	Improved RBNN		
			Linear Kernel	Polynomial Kernel	Quadratic Kernel
Ling spam	35	25	7	9	8
Enron	31	23	5	7	6
Spam Assassin	37	29	12	12	13
CSDM C2010	32	21	7	11	9
PU Dataset	39	30	12	15	14

The following Figure 3 shows the graphical representation of the proposed method for execution time which proves that the proposed method has very less execution time when comparing the RBNN with SVM method. Improved RBNN is applied to get the final best subset for SVM classifier and overcomes local optima problem using dynamic adaptation approach by dynamically changing inertia weight value. Too small parameter adjustment would cause too small particle movement and results in useful data, but it is time consuming. In each iteration, the global best particle is being chosen according to their fitness values. Since the global best particles in the improved RBNN are selected with the least local best particles. Therefore, optimal feature space is identified with a lesser amount of time. The execution time for the improved RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 8, 9, and 7 seconds for Ling spam dataset whereas RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 12, 13, and 11 seconds for Ling spam dataset. The execution time for the improved RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 6, 7, and 5 seconds for Enron spam dataset whereas RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 10, 12, and 9 seconds for Enron spam dataset. The execution time for the improved RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 13, 12, and 10 seconds for Spam Assassin dataset whereas RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 16, 17, and 15 seconds for Spam Assassin dataset. The execution time for the improved RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 9, 11, and 7 seconds for CSDMC2010 spam dataset whereas RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 17, 15, and 14 seconds for CSDMC2010 spam dataset. The execution time for the improved RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 14, 15, and 12 seconds for PU spam dataset whereas RBNN with quadratic kernel SVM, polynomial kernel SVM, linear kernel SVM is 20, 19, and 18 seconds for PU spam dataset.

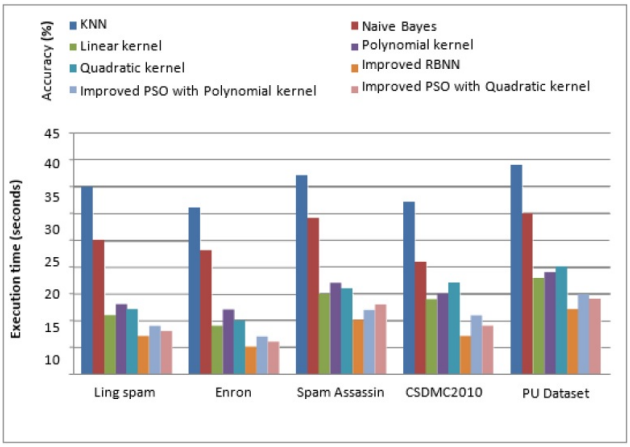

Figure 3. Execution Time using Improved RBNN

C. Precision and Recall (%)

Table 3 and 4 shows the precision and recall values obtained from KNN, Naive Bayes, and SVM classifiers.

TABLE III. Precision (%) using Improved RBNN

Data Sets	KNN	Naive Bayes	Improved RBNN		
			Linear Kernel	Polynomial Kernel	Quadratic Kernel
Ling spam	83.12	88.25	98.72	96.57	95.66
Enron	85.89	90.01	96.54	93.12	94.78
Spam Assassin	81.64	99.31	96.87	92.66	97.69
CSDM C2010	88.58	91.98	94.22	93.58	91.27
PU Dataset	84.45	89.57	99.14	92.36	95.35

TABLE IV. Recall (%) using Improved RBNN

Data Sets	KNN	Naive Bayes	Improved RBNN			
			Linear Kernel	Linear Kernel	Polynomial Kernel	Quadratic Kernel
Ling spam	35.97	31.81	19.74	7.29	10.57	12.65
Enron	38.06	36.75	16.54	8.21	11.34	18.22
Spam Assassin	37.86	34.38	21.18	9.87	13.35	17.56
CSDM C2010	41.73	37.11	20.31	9.84	15.64	13.87
PU Dataset	40.12	39.86	29.88	15.37	17.00	19.63

The graphical representation of the proposed method for recall value which proves that the proposed method attained less recall value using SVM classifier. The recall value for the improved RBNN and RBNN with quadratic kernel SVM is 12.65% and 26.44%, recall value for improved RBNN and RBNN with polynomial kernel SVM is 10.54% and 22.12%, recall value for improved RBNN and RBNN with linear kernel SVM is 7.29% and 19.74% for Ling spam dataset. The recall value for the improved RBNN and RBNN with quadratic kernel SVM is 18.22% and 31.54%, recall value for improved RBNN and RBNN with polynomial kernel SVM is 11.34% and 25.73%, recall value for improved RBNN and RBNN with linear kernel SVM is 8.21% and 16.54% for Enron spam dataset. The recall value for the improved RBNN and RBNN with quadratic kernel SVM is 17.56% and 36.36%, recall value for improved RBNN and RBNN with polynomial kernel SVM is 13.35% and 26.64%, recall value for improved RBNN and RBNN with linear kernel SVM is 9.87% and 21.18% for Spam Assassin dataset. The recall value for the improved RBNN and RBNN with quadratic kernel SVM is 13.87% and 33.78%, recall value for improved RBNN and RBNN with polynomial kernel SVM is 15.64% and 28.94%, recall value for improved RBNN and RBNN with linear kernel

SVM is 9.48% and 20.31% for CSDMC2010 spam dataset. The recall value for the improved RBNN and RBNN with quadratic kernel SVM is 19.63% and 39.94%, recall value for improved RBNN and RBNN with polynomial kernel SVM is 17.00% and 33.66%, recall value for improved RBNN and RBNN with linear kernel SVM is 15.37% and 29.88% for PU spam dataset.

TABLE V. Error Rate using Improved RBNN

Data Sets	KNN	Naive Bayes	Improved RBNN		
			Linear Kernel	Polynomial Kernel	Quadratic Kernel
Ling spam	0.2468	0.2252	0.0426	0.0613	0.0622
Enron	0.2616	0.2488	0.0535	0.0694	0.0705
Spam Assassin	0.1923	0.1465	0.0521	0.0734	0.0781
CSDM C2010	0.2376	0.2135	0.0703	0.0802	0.0810
PU Dataset	0.2446	0.1902	0.0346	0.0635	0.0452

The graphical representation of the proposed method for the error rate which proves that the proposed method achieved less error rate using the SVM classifier. Large inertia weight enhances global search while small inertia weight results in faster convergence. The best particle is being chosen according to their fitness values and the number of particles selected is also updated if it has a reduced value. The velocity of the particle in each dimension increases rapidly and tends to select a higher number of particles for obtaining high classification accuracy which drastically reduces the error rate.

5. CONCLUSION

In this study, we have introduced a feature selection method based on improved RBNN for spam classification. Here, SVM performs well for spam classification problems. The results obtained show that an improved RBNN approach gives a better classification in terms of accuracy. RBNN easily suffers from partial optimism which leads to premature convergence. It also requires large amounts of memory, which may limit its implementation in resource-poor areas.

REFERENCES

[1] H. Hou, Y. Chen, R. Beyah, and Y.-Q. Zhang, "Filtering spam by using factors hyperbolic tree," in *IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference*. IEEE, 2008, pp. 1–5.

[2] T. M. Maina, "Instant messaging an effective way of communication in workplace," *arXiv preprint arXiv:1310.8489*, 2013.

[3] A. M. Alhawamleh and A. Ngah, "Knowledge sharing among jordanian academicians: A case study of tafila technical university (ttu) and mutah university (mu)," in *2017 8th International Conference on Information Technology (ICIT)*. IEEE, 2017, pp. 262–270.

[4] O. Kufandirimbwa and R. Gotora, "Spam detection using artificial neural networks (perceptron learning rule)," *Online Journal of Physical and Environmental Science Research*, vol. 1, no. 2, pp. 22–29, 2012.

[5] G. L. Wittel and S. F. Wu, "On attacking statistical spam filters." in *CEAS*, 2004.

[6] R. Anderson, C. Barton, R. Böhme, R. Clayton, M. J. Van Eeten, M. Levi, T. Moore, and S. Savage, "Measuring the cost of cybercrime," in *The economics of information security and privacy*. Springer, 2013, pp. 265–300.

[7] C. Daehnick, I. Klinghoffer, B. Maritz, and B. Wiseman, "Large leo satellite constellations: Will it be different this time," *McKinsey and Company*, vol. 4, 2020.

[8] M. Islam and M. Chowdhury, "Spam filtering using ml algorithms," 2005.

[9] N. Zhang, Y. Jiang, B. Fang, X. Cheng, and L. Guo, "Traffic classification-based spam filter," in *2006 IEEE International Conference on Communications*, vol. 5. IEEE, 2006, pp. 2130–2135.

[10] M. A. Mohammed, D. A. Ibrahim, and A. O. Salman, "Adaptive intelligent learning approach based on visual anti-spam email model for multi-natural language," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 774–792, 2021.

[11] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, p. 89, 2019.

[12] B. L. Ott and R. L. Mack, *Critical media studies: An introduction*. John Wiley & Sons, 2020.

[13] S. Youn and D. McLeod, "A comparative study for email classification," in *Advances and innovations in systems, computing sciences and software engineering*. Springer, 2007, pp. 387–391.

[14] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.

[15] A. Çıltık and T. Güngör, "Time-efficient spam e-mail filtering using n-gram models," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 19–33, 2008.

[16] L. Enciso, R. Baez, A. Maldonado, E. Zelaya-Policarpo, and P. A. Quezada-Sarmiento, "E-mail client multiplatform for the transfer of information using the smtp java protocol without access to a browser," in *World Conference on Information Systems and Technologies*. Springer, 2018, pp. 1124–1130.

[17] A. G. Lopez-Herrera, E. Herrera-Viedma, and F. Herrera, "A multiobjective evolutionary algorithm for spam e-mail filtering," in *2008 3rd International Conference on Intelligent System and Knowledge Engineering*, vol. 1. IEEE, 2008, pp. 366–371.

[18] M. Baykara and Z. Z. Gürel, "Detection of phishing attacks," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*. IEEE, 2018, pp. 1–5.

[19] A. Hawamleh, A. S. M. Alorfi, J. A. Al-Gasawneh, and G. Al-Rawashdeh, "Cyber security and ethical hacking: The importance of protecting user data," *Solid State Technology*, vol. 63, no. 5, pp. 7894–7899, 2020.

[20] L. Pei-yu, Z. Li-wei, and Z. Zhen-fang, "Research on e-mail filtering based on improved bayesian," *Journal of Computers*, vol. 4, no. 3, pp. 271–275, 2009.

[21] H. M. Hirei, "Investigating and validating scam triggers: A case study of a craigslist website," 2020.

[22] A. M. K. Alhawamleh, "Web based english placement test system (elpts)," Ph.D. dissertation, Universiti Utara Malaysia, 2012.

[23] A. Kumari, S. Tanwar, S. Tyagi, and N. Kumar, "Verification and validation techniques for streaming big data analytics in internet of things environment," *IET Networks*, vol. 8, no. 3, pp. 155–163, 2019.

[24] A. M. El-Halees, "Filtering spam e-mail from mixed arabic and english messages: A comparison of machine learning techniques." *International Arab Journal of Information Technology (IAJIT)*, vol. 6, no. 1, 2009.

[25] M. Vijayalakshmi, S. Mercy Shalinie, M. H. Yang, and R. M. U, "Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions," *Iet Networks*, vol. 9, no. 5, pp. 235–246, 2020.

[26] J. Göbel, T. Holz, and P. Trinius, "Towards proactive spam filtering," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2009, pp. 38–47.

[27] Z. Pan, S. Hariri, and J. Pacheco, "Context aware intrusion detection for building automation systems," *Computers & Security*, vol. 85, pp. 181–201, 2019.

[28] S. Rao, A. K. Verma, and T. Bhatia, "A review on social spam detection: challenges, open issues, and future directions," *Expert Systems with Applications*, vol. 186, p. 115742, 2021.

[29] W. H. Dutton, *Social transformation in an information society: Rethinking access to you and the world*. Unesco Paris, 2004, vol. 13.

[30] J. R. Levine, "Experiences with greylisting." in *CEAS*, 2005.

[31] M. Basavaraju and D. R. Prabhakar, "A novel method of spam mail detection using text based clustering approach," *International Journal of Computer Applications*, vol. 5, no. 4, pp. 15–25, 2010.

[32] A. N. Pour, R. Kholghi, and S. B. Roudsari, "Minimizing the time of spam mail detection by relocating filtering system to the sender mail server," *arXiv preprint arXiv:1208.5556*, 2012.

[33] M. A. Siddiqui, *Data mining methods for malware detection*. University of Central Florida, 2008.

[34] M. Zareapoor, K. Seeja, and M. A. Alam, "Analysis on credit card fraud detection techniques: based on certain design criteria," *International journal of computer applications*, vol. 52, no. 3, 2012.

[35] J. Al-Gasawneh, A. AL-Hawamleh, A. Alorfi, and G. Al-Rawashde, "Moderating the role of the perceived security and endorsement on the relationship between per-ceived risk and intention to use the artificial intelligence in financial services," *International Journal of Data and Network Science*, vol. 6, no. 3, pp. 743–752, 2022.

[36] D. Kalbande, H. Panchal, N. Swaminathan, and P. Ramaraj, "Anfis based spam filtering model for social networking websites," *International Journal of Computer Applications*, vol. 44, no. 11, pp. 0975–8887, 2012.

[37] M. Bassiouni, M. Ali, and E. El-Dahshan, "Ham and spam e-mails classification using machine learning techniques," *Journal of Applied Security Research*, vol. 13, no. 3, pp. 315–331, 2018.

[38] C. O'Brien and C. Vogel, "Spam filters: Bayes vs. chi-squared; letters vs. words," *ISICT*, vol. 3, pp. 291–296, 2003.

[39] T. B. Shahi, A. Yadav *et al.*, "Mobile sms spam filtering for nepali text using naïve bayesian and support vector machine," *International Journal of Intelligence Science*, vol. 4, no. 01, pp. 24–28, 2014.

[40] P. Sudhakar, G. Poonkuzhali, K. Thiagarajan, R. K. Keshav, and K. Sarukesi, "Fuzzy logic for e-mail spam deduction," in *Proceedings of the 10th WSEAS international conference on Applied computer and applied computational science*, 2011, pp. 83–88.

[41] H. T. Elshoush and E. A. Dinar, "Using adaboost and stochastic gradient descent (sgd) algorithms with r and orange software for filtering e-mail spam," in *2019 11th Computer Science and Electronic Engineering (CEEC)*. IEEE, 2019, pp. 41–46.

[42] A. N. Soni, "Spam-e-mail-detection-using-advanced-deep-convolution-neuralnetwork-algorithms," *Journal for innovative development in pharmaceutical and technical science*, vol. 2, no. 5, pp. 74–80, 2019.

[43] A. H. Mohammad and R. A. Zitar, "Application of genetic optimized artificial immune system and neural networks in spam detection," *Applied Soft Computing*, vol. 11, no. 4, pp. 3827–3845, 2011.

[44] A. Khorsi, "An overview of content-based spam filtering techniques," *Informatica*, vol. 31, no. 3, 2007.

[45] I. Saini, D. Singh, and A. Khosla, "Qrs detection using k-nearest neighbor algorithm (knn) and evaluation on standard ecg databases," *Journal of advanced research*, vol. 4, no. 4, pp. 331–344, 2013.

[46] H. Bhavsar and M. H. Panchal, "A review on support vector machine for data classification," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 1, no. 10, pp. 185–189, 2012.

[47] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[48] G. Santhi, S. M. Wenisch, and P. Sengutuvan, "Fuzzy rule based novel approach to spam filtering," *International Journal of Computer Applications*, vol. 71, no. 14, 2013.

[49] D. Sonia, "Spam filter: Vsm based intelligent fuzzy decision maker," *International Journal of Computer Science and Technology*, vol. 1, no. 1, pp. 48–52, 2010.

[50] J. N. Shrivastava and M. H. Bindu, "Trends, issues and challenges concerning spam mails," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 4, no. 8, p. 10, 2012.

[51] M. M. Fuad, D. Deb, and M. S. Hossain, "A trainable fuzzy spam detection system," in *Proc. of the 7th Int. Conf. on Computer and Information Technology*, 2004.

[52] M. Yeganeh, L. Bin, and G. Babu, "A model for fuzzy logic based machine learning approach for spam filtering," *IOSR J. Computer Engineering*, vol. 4, pp. 07–10, 2012.

[53] M. Begol and K. Maghooli, "Improving digital image edge detection by fuzzy systems," *World Academy of Science, Engineering and Technology*, vol. 81, pp. 76–79, 2011.

[54] G. Goertz and J. Mahoney, "Concepts and measurement: Ontology

and epistemology," *Social Science Information*, vol. 51, no. 2, pp. 205–216, 2012.

[55] V. Christina, S. Karpagavalli, and G. Suganya, "A study on email spam filtering techniques," *International Journal of Computer Applications*, vol. 12, no. 1, pp. 0975–8887, 2010.

[56] E.-S. M. El-Alfy and F. S. Al-Qunaieer, "A fuzzy similarity approach for automated spam filtering," in *2008 IEEE/ACS International Conference on Computer Systems and Applications*. IEEE, 2008, pp. 544–550.

[57] A. Kanaan, A. AL-Hawamleh, A. Abulfaraj, H. Al-Kaseasbeh, and A. Alorfi, "The effect of quality, security and privacy factors on trust and intention to use e-government services," *International Journal of Data and Network Science*, vol. 7, no. 1, pp. 185–198, 2023.

Ahmad Mtair AL-Hawamleh Institute of Public Administration AL-Hawamleh is Assistant Professor in Computer Science-Cybersecurity at the Institute of Public Administration-KSA, Certified Trainer in Blackboard Education Technology & Services, and Certified Trainer in Zoom Meetings Platform. His research is situated in the field of Information Security, Cybersecurity, and IoT. AL-Hawamleh has worked as a lecturer at the faculty of Engineering and Computer Science at Tafila Technical University-Jordan, lecturer at the faculty of Computer Science at the Saudi Electronic University-KSA, and he is currently working as an Assistant Professor at the Electronic Training Department at the Institute of Public Administration-KSA.