I. J. Computer Network and Information Security, 2017, 7, 45-58
Published Online July 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2017.07.06

K-MLP Based Classifier for Discernment of Gratuitous Mails using N-Gram Filtration

Harjot Kaur
CT Group of Institution/CSE, Jalandhar, 144041, India
E-mail: harjotkaur844@gmail.com

Er. Prince Verma
CT Group of Institution/CSE, Jalandhar, 144041, India
E-mail: prince.researchwork@gmail.com

***Abstract*—**Electronic spam is a highly concerning phenomenon over the internet affecting various organisations like Google, Yahoo etc. Email spam causes several serious problems like high utilisation of memory space, financial loss, degradation of computation speed and power, and several threats to authenticated account holders. Email spam allows the spammers to deceit as a legitimate account holder of the organisations to fraud money and other useful information from the victims. It is necessary to control the spreading of spam and to develop an effective and efficient mechanism for defence. In this research, we proposed an efficient method for characterising spam emails using both supervised and unsupervised approaches by boosting the algorithm's performance. This study refined a supervised approach, MLP using a fast and efficient unsupervised approach, K-Means for the detection of spam emails by selecting best features using N-Gram technique. The proposed system shows high accuracy with a low error rate in contrast to the existing technique. The system also shows a reduction in vague information when MLP was combined with K-Means algorithm for selecting initial clusters. N-Gram produces 100 best features from the group of data. Finally, the results are demonstrated and the output of the proposed technique is examined in contrast to the existing technique.

***Index Terms*—**E-Mail, Spam Filters, N-Gram feature selection, K-Means clustering algorithm, Multi-Layer Perceptron Neural Network (MLP-NN) algorithm, Support Vector Machine (SVM) algorithm.

I. INTRODUCTION

Email is the most efficient and fastest mode of communication to exchange information over the internet. Due to the increase in the number of account holders over the various social sites, there is a tremendous increase in the rate of spreading of spam emails. Despite having various tools available still, there are many sources for the spam to originate. Lack of defence mechanism to prevent the spreading of spam can cause severe economic loss, loss of bandwidth for handling spam emails, memory utilisation and can cause personal and monetary threats to the information holders. Spam can be understood as 'an unwanted illegitimate, junk emails received by the legitimate users from unauthenticated sources'. To handle spam emails, spam filtration technique is followed which blocks the spam email from entering into the mail inbox, but the major issue with spam filtration is that a valid email can be detected as spam or a spam email can be missed. Spam can be filtered by a non-machine learning and machine learning techniques. Some of the non-machine learning technique used to filter spam emails are black list/white list, signatures, email header analysis [10]. A black list is a technique that sorts the addresses of the contacts that are unknown to the recipients that may contain spam emails while a white list is the list of known contacts to the recipients. The signature-based technique detects the spam using a hash value for comparison with new emails received while header analysis of emails involves a set of rules to detect if an email is a spam or ham. Machine learning technique shows better classification results than non-machine learning techniques and they are categorised into supervised and unsupervised approach viz. K-Means, SVM, MLP, Decision tree based classification [25].

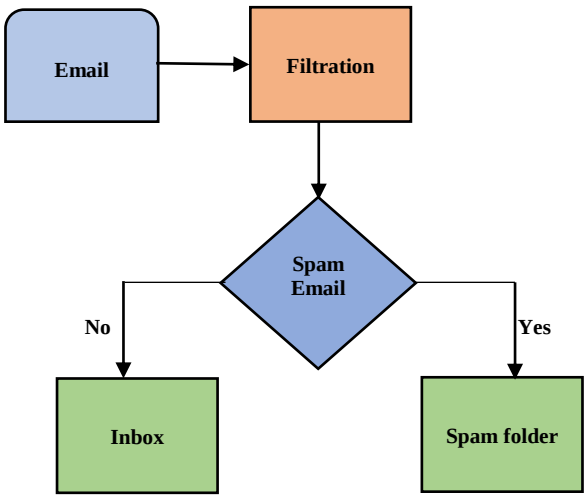


Fig.1. Email Spam Detection Identification Process.

Emails can be categorised as a spam email if it shows

following characteristics [53]:

1. *Unsolicited Email:* Email received from unknown contact or illegitimate contact.
2. *Bulk Mailing:* The type of email which is sent in bulk to many users.
3. *Nameless Mails:* The type of emails in which the identity of the user is not shown or is hidden.

The success ratio of machine learning algorithms over non-machine learning algorithms is more. These techniques work by selecting the best features from the data to group the emails as spam or ham. Feature selection can be carried out in two ways [53]:

1. *Header Based Selection:* Selecting the best feature from the header of the mail. It contains sender's address, BCC (Blind Carbon Copy), CC (Carbon Copy), To, From, Date and Subject.
2. *Content Based Selection:* Selecting the best feature from the content in the mail. It contains the main message either in the form of text, audio or video, attachments etc.

Content Based Feature Selection is proven as the most authenticated feature selection as compared to Header Based as Header Based Feature Selection can be easily tempered by the hackers or spammers [53].

The paper is organised as follows, Section 2 defines the related work carried till date, Section 3 defines the algorithms undertaken for the research, and Section 4 provides our proposed framework for email spam detection. Section 5 shows the results and its discussion and Section 6 concludes our proposed methodology in comparison to the existing methodology.

II. Related Work

This section describes various papers related to the work carried on detection of spam emails.

Bo Yu and Zong-ben Xu (2008) performed a comparative analysis on content-based spam classification using different machine learning algorithms. This paper classified spam emails using four different machine learning algorithms viz. Naıve Bayesian, Neural Network, Support Vector Machine and Relevance Vector Machine. The analysis was performed on the different training dataset and feature selection. Analysis results demonstrated that NN algorithm is not a good algorithm to be used as a tool for spam rejection. SVM and RVM machine learning algorithms are better algorithms than NB classifier. Instead of slow learning, RVM is still better algorithm than SVM for spam classification with less execution time and fewer relevance vectors [1].

Tiago A. Almeida and Akebo Yamakami (2010) performed a comparative analysis using content-based filtering for spam. This paper discussed seven different modified versions of Naïve Bayes Classifier and compared those results with Linear Support Vector Machine on six different open and large datasets. The results demonstrated that SVM, Boolean NB and Basic NB are the best algorithms for spam detection. However, SVM executed the accuracy rate higher than 90% for almost all the datasets utilised [2].

Loredana Firte, Camelia Lemnaru and Rodica Potolea (2010) performed a comparative analysis on spam detection filter using KNN Algorithm and Resampling approach. This paper makes use of the K-NN algorithm for classification of spam emails on the predefined dataset using feature's selected from the content and emails properties. Resampling of the datasets to appropriate set and positive distribution was carried out to make the algorithm efficient for feature selection [3].

Ms.D.Karthika Renuka, Dr.T.Hamsapriya, et. al. (2011) performed a comparative analysis of spam classification based on supervised learning using several machine learning techniques. In this analysis, the comparison was done using three different machine learning classification algorithms viz. Naïve Bayes, J48 and Multilayer perceptron (MLP) classifier. Results demonstrated high accuracy for MLP but high time consumption. While Naïve Bayes accuracy was low than MLP but was fast enough in execution and learning. The accuracy of Naïve Bayes was enhanced using FBL feature selection and used filtered Bayesian Learning with Naïve Bayes. The modified Naïve Bayes showed the accuracy of 91% [4].

Rushdi Shams and Robert E. Mercer (2013) performed a comparative analysis of the classification of spam emails by using text and readability features. This paper proposed an efficient spam classification method along with feature selection using the content of emails and readability. This paper used four data sets such as CSDMC2010, Spam Assassin, Ling-spam, and Enron-spam. Features are categorised into three categories i.e. traditional features, test features and readability features. The proposed approach is able to classify emails of any language because the features are kept independent of the languages. This paper used five classification based algorithms for spam detection viz. Random Forest (RF), Bagging, Adaboostm 1, Support Vector Machine (SVM) and Naïve Bayes (NB). Results comparison among different classifiers predicted Bagging algorithm to be the best for spam detection [5].

Megha Rathi and Vikas Pareek(2013) performed an analysis on spam email detection through Data Mining by performing analysis on classifiers by selecting and without selecting the features [6].

Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta and Anuja Arora (2014) performed a comparative analysis of text and images by using KNN, Naïve Bayes and Reverse-DBSCAN Algorithm for email spam detection. This analysis paper proposed a methodology for detecting text and spam emails. They used Naïve Bayes, K-NN and a modified Reverse DBSCAN (Density- Based Spatial Clustering of Application with Noise) algorithm. Authors used Enron dataset for text and image spam classification. They used Google's open source library, Tesseract for extracting words from images. Results show that these three machine learning

algorithms give better results without pre-processing among which Naïve Bayes algorithm is highly accurate than other algorithms [7].

Savita Pundalik Teli and Santosh Kumar Biradar (2014) performed an analysis of effective email classification for spam and non-spam emails [8].

Izzat Alsmadi and Ikdam Alhami (2015) performed an analysis on clustering and classification of email contents for the detection of spam. This paper collected a large data set of personal emails for the spam detection of emails based on folder and subject classification. Supervised approach viz. classification alongside unsupervised approach viz. clustering was performed on the personal data set. This paper used SVM classification algorithm for classifying the data obtained from K-means clustering algorithm. This paper performed three types of classification viz. without removing stop words, removing stop words and using N-gram based classification. The results clearly illustrated that N-gram based classification for spam detection is the best approach for large and Bi-language text [9].

Ali Shafigh Aski and Navid Khalilzadeh Sourati (2016), filtration was carried out using machine learning algorithms namely, Naïve Bayes, J48 and MLP on the personal data set collected during a six-month period including 750 spam and 750 ham emails. Results analysis showed that MLP has higher accuracy of 99.3% than other two algorithms but the computational time for spam detection in MLP was high 138.05 sec. than other algorithms. In the same research, Naïve Bayes shows a slightly low accuracy of 98.6 % than MLP but the computational time of Naïve Bayes was low than MLP by 0.15 seconds [10].

III. PRELIMINARIES

In this section, the author presents the algorithms undertaken for the research work. The detection of the email spamming is conducted by first performing the filtration by N-Gram based filtration, then clustering by K-Means is performed on the email dataset for specifying two base clusters viz. spam and ham clusters. In the last step, classification of the clustered data is performed by MLP Neural Network for validating the clustering results and labelling the emails to two defined classes ham and spam classes. The results are compared with the existing N-Gram-K-SVM technique for various parameters viz. Accuracy, Sensitivity, Specificity, F-Measure, Precision, and Root mean square error. Initially the comparison is performed on Simple MLP and SVM after pre-processing the raw data, later on, the results are compared by implementing the N-Gram based filtration for SVM and MLP by comparing Bi-gram, Tri-gram, and Four-gram, and finally in the last step our proposed approach of refining the MLP with K-Means algorithm by using N-gram based feature selection technique is compared with the existing N-Gram-K-SVM technique. The advantages of MLP algorithm like generalisation and highly fault tolerant makes it an efficient algorithm over the SVM algorithm.

A. K-Means Cluster Analysis Algorithm

Clustering is an unsupervised approach for splitting the collection of data items into several clusters. The partitioning of the data items is carried out by the maximum similarity measure. K-Means is the most widely used algorithm that efficiently assigns the data objects in a cluster. K-Means works by grouping the information objects into 'k' clusters. Let dataset 'S', contain 'n' number of information objects where 'k' is the number of clusters formed, K-Means algorithm assign 'n' number of information objects to 'k' clusters, where ($k \leq n$). Algorithmic steps for K-Means algorithm are as below:

1. Manually nominate the cluster centre "c".
2. Observe the distance between every information point and the selected centre of the clusters.
3. Label the information point to the cluster centre whose detachment is the minimum from the cluster centre.
4. Recalculate the new group centre by equation number 1:

$$c_i{}^* = (1/n_i)\sum s_j^i \text{ for i=1,2,….,k.} \quad (1)$$

Where "n_i" represent the number of information points in the i^{th} cluster.

5. Recalculate the distance between every information point and newly formulated cluster centre.
6. After achieving convergence stop the algorithm else repeat from step 3.

B. Multi-Layer Perceptron Neural Network

Multi-Layer Perceptron (MLP) is a feed-forward data processing network that maps the group of inputs to their corresponding outputs. Fig. 1 demonstrates a feed-forward multilayer perceptron neural network [10], [18], [25]. MLP is made up of simple neurons termed as perceptron's and is similar to the human nervous system. Neural network generates information by enabling input perceptron's consisting the values labelled on them. MLP intakes multiple layers and every corresponding layer are connected to the next single layer with weights specified on them. The activation function of neurons is calculated by the formula mentioned below for generating the output of the layer [10], [25]:

$$a_i = \sigma(\sum_j W_{ij}\, O_j) \quad (2)$$

Where a_i represent the level of activation for i^{th} neurons; j is the set of neurons of the previous layer; W_{ij} is the weight of the connection between neurons i and j; O_j represents the output of j^{th} neuron and $\sigma(x)$ is the transfer function. For algorithm refer paper [53].

$$\sigma(x) = \frac{1}{1+e^x} \quad (3)$$

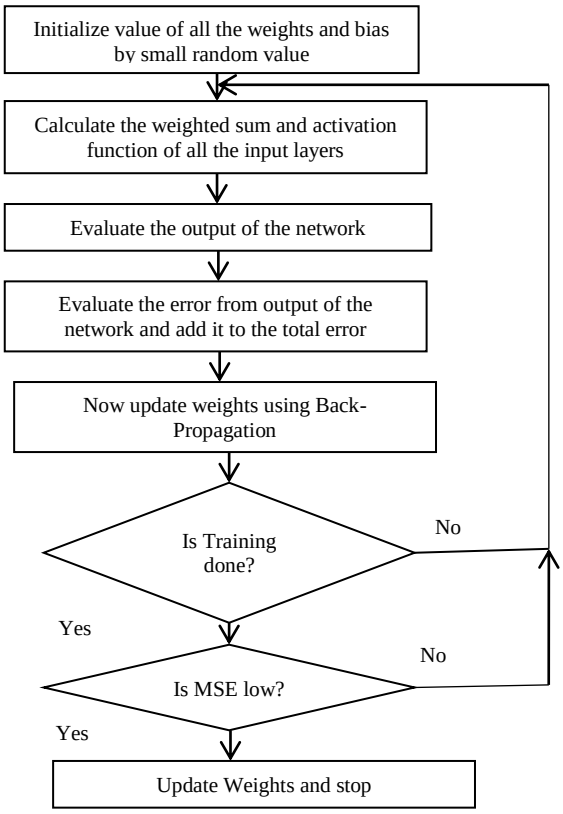


Fig.2. Layer Perceptron Process Diagram.

C. N-Gram based Feature Selection

Feature selection is a measure of choosing the most optimal features from the dataset for better classification results from the pre-processed dataset. In this research, N-Gram based feature selection is used which is a predictive algorithm used for predicting the probability of the outcome of next word after making observations for *N-1* words in a sentence or text corpus. N-Gram has its application in text mining and natural language processing. N-grams are the set of co-occurring words that move one or *X* (number of words in a corpus) steps ahead while executing N-Grams [25].

Let *X*, be the number of words in a given text corpus *T*, the number of N-Grams can be calculated by:

$$N_{grams}(T) = X - (N-1) \quad (4)$$

N-Grams varies in size where N= 1, 2, 3 and so on. In the research result analysis was carried for n=4, representing the size of N-Grams to avoid the formation of the complete sentence [25].

1. *Uni-Gram:* The N-Gram in which the size of 'n' is one is termed as Uni-Gram. For example, the word "GOOD" in Uni-Gram can be processed by moving one step ahead viz. "G to O", "O to O", "O to D"
2. *Bi-Gram:* The N-Gram in which the size of 'n' is two is termed as bi-Gram. For example, "GOOD" in Bi-Gram can be processed by moving two steps ahead in the string of data viz. "GO to OO", "OO to OD".
3. *Tri-Gram:* The N-Gram in which the size of 'n' is three is termed as Tri-Gram and so on for N= 4, N=5 etc.

N-Gram for a text corpus "Boys were playing football on the ground" using Bi-Gram (N=2) will be "Boys were", "were playing", "playing football", "football on", "on the", "the ground". In the example of word corpus containing 7 N-grams is illustrated as we move two steps ahead for generating the possibility of occurrence of next word.

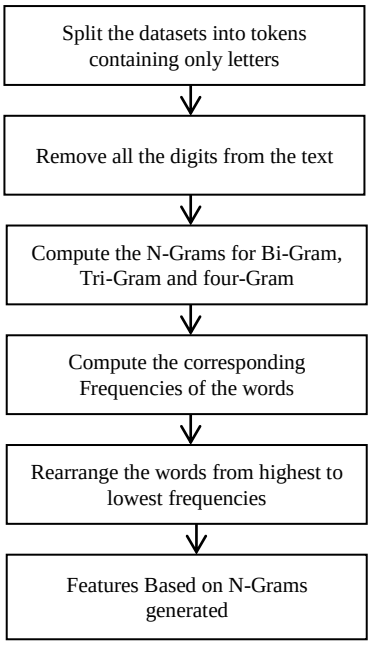


Fig.3. N-Gram Based Feature Selection Technique Process Model.

D. Support Vector Machine

Support vector machine (SVM) is a supervised approach for machine learning. The main idea used in SVM is constructing a hyperplane that is optimal for the classification of patterns that can be linearly separated [53]. This algorithm work by plotting each information point in the n-dimensional workspace, where n represents the number of features which are equal to the coordinates in the workspace. The optimal hyperplane differentiates the classes at this point [53].

In email spam detection, the aim is to divide the email into two groups, spam or ham email by using an optimal hyperplane. The idea is to distinguish the two classes to achieve maximum marginal difference between two classes, viz. spam and ham. SVM represents the information points in the workspace, mapped so that the information points of the other groups are partitioned by a maximum marginal difference. New information points are labelled to that same workspace and predictions are conducted to analyse the category of the new information point. SVM can efficiently perform non-linear classification by kernel trick (similarity function) [53].

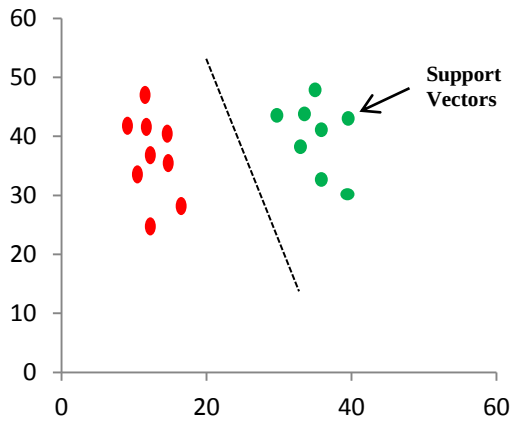


Fig.4. SVM Representing The Difference between Two Classes Using Hyper-Plane [53].

Algorithmic steps of SVM for the classification process are as follows [53].

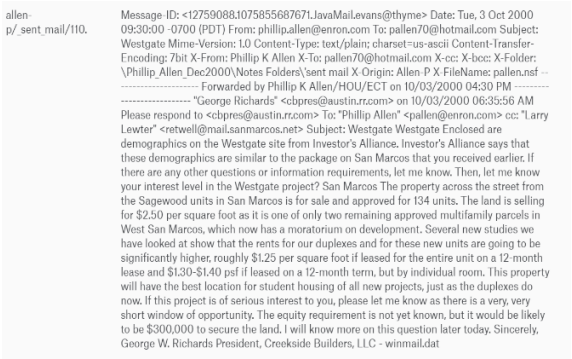
1. Train the initial SVM using all the training data to have support vectors decision functions.
2. Eliminate those support vectors generated from the training of initial SVM whose projections have greatest curvatures on the hypersurface by: finding the projection of the support vectors along the gradient of decision function used, calculate the notion of curvature for every support vector on the hyperplane, lastly sort the support vectors in the decreasing order and deduct the top N-percentage of the vectors of support.
3. Retrain the SVM by left over vectors for best decision.
4. Use the group of information point to finally train the SVM, generating support vectors.

IV. Framework Of The Proposed Methodology For The Detection Of Spam

Email is primarily the most common method of communication over the internet and emails can be categorised as spam or ham. Spam emails are sent to the recipients in bulk and are unwanted to account holders. These types of spam emails are very serious and it is, therefore, important to manage the emails and raising problems of misuse between people and organisations. A major requirement is to protect the authenticated account holders from the spam emails. In this section, the author presents a proposed approach for discernment of gratuitous emails. The proposed methodology comprises of various steps: (1) Dataset pre-processing (2) Feature Selection using N-Gram (3) Cluster analysis by K-Means (4) Classification by MLP. The comparison of the proposed technique is carried out with the existing approach which uses SVM algorithm for classification of spam emails. The results are conducted on Enron data set by reducing the features for better analysis.

A. Dataset

For the implementation of proposed methodology, Enron dataset in arff (attribute relation file format) format is used which comprises of 5 lakh personal emails of 150 employees of the Enron Corporation collected from UCI Resource Repository. In the research work, 100 best features are used with 50% of spam rate and 50% ham rate.


allen-p/_sent_mail/110.

Message-ID: <12759088.1075855687671.JavaMail.evans@thyme> Date: Tue, 3 Oct 2000 09:30:00 -0700 (PDT) From: phillip.allen@enron.com To: pallen70@hotmail.com Subject: Westgate Mime-Version: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Phillip K Allen X-To: pallen70@hotmail.com X-cc: X-bcc: X-Folder: \Phillip_Allen_Dec2000\Notes Folders\'sent mail X-Origin: Allen-P X-FileName: pallen.nsf

---------- Forwarded by Phillip K Allen/HOU/ECT on 10/03/2000 04:30 PM ----------

"George Richards" <cbpres@austin.rr.com> on 10/03/2000 06:35:56 AM

Please respond to <cbpres@austin.rr.com> To: "Phillip Allen" <pallen@enron.com> cc: "Larry Lewter" <retwell@mail.sanmarcos.net> Subject: Westgate Westgate Enclosed are demographics on the Westgate site from Investor's Alliance. Investor's Alliance says that these demographics are similar to the package on San Marcos that you received earlier. If there are any other questions or information requirements, let me know. Then, let me know your interest level in the Westgate project? San Marcos The property across the street from the Sagewood units in San Marcos is for sale and approved for 134 units. The land is selling for $2.50 per square foot as it is one of only two remaining approved multifamily parcels in West San Marcos, which now has a moratorium on development. Several new studies we have looked at show that the rents for our duplexes and for these new units are going to be significantly higher, roughly $1.25 per square foot if leased for the entire unit on a 12-month lease and $1.30-$1.40 psf if leased on a 12-month term, but by individual room. This property will have the best location for student housing of all new projects, just as the duplexes do now. If this project is of serious interest to you, please let me know as there is a very, very short window of opportunity. The equity requirement is not yet known, but it would be likely to be $300,000 to secure the land. I will know more on this question later today. Sincerely, George W. Richards President, Creekside Builders, LLC - winmail.dat

Fig.5. Text Message Received by an Employee.

B. Pre-Processing

To pre-process the Enron dataset and to extract useful emails for the detection of spam, three pre-processing techniques are followed.

1. *Lexical Analysis and Tokenization:* Firstly, Lexical Analysis and Tokenization of the text document is performed by which the text is split into words of individual identity to create a bag of words model by using string to word vector filter. The collection of email dataset is represented as a vector of the matrix representing the rows of $m * n$ order, where m represents the text corpus and n is the list of features. The reason to split the sentence corpus into words is to avoid the performance degradation of the algorithms on the large datasets.
2. *Stop Words Removal:* In the second step, stop words viz. of, and, the, etc. are removed by using stop word filter also the words with the highest frequency of occurrence by computing the TF-IDF (Term Frequency - Inverse document frequency) of the words in the dataset are removed. The frequency of the words ranges from 1 to 100,000 times. In the research work, the words with 100 frequency value and above are used, while inverse document frequency (IDF) calculated the importance of the word.
3. *Stemming:* In the third step of pre-processing, stemming from the dataset is carried out. Stemming removed the derived words like connect, connecting and connected that has the same meaning. By performing pre-processing of the dataset, the dimensionality of the dataset is reduced. Dimensionality reduction technique reduces the number of features in the set of data. The main reason to reduce the features of the dataset is that most of the algorithms slow down due to the unwanted features. In this research methodology, we reduced the dimensionality of

the features by using all the above techniques. Finally, 100 best features are selected using N-Gram based feature selection technique.

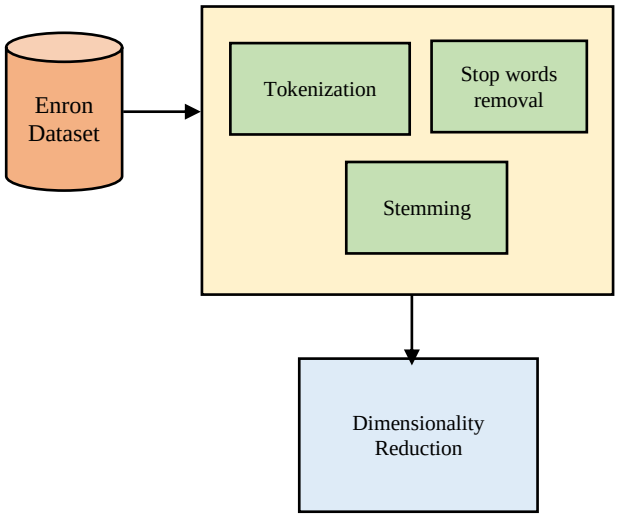


Fig.6. Process for the Dimensionality Reduction.

C. N-Gram based Feature Selection

In the research work, 100 best features are selected from the data corpus using N-Gram based feature selection technique. In the research work, comparisons are performed for bi-gram, tri-gram, and four-gram. It was observed that with the increase in the words comparison for N=5 (five words) the performance of the algorithms starts decreasing due to the formation of sentence corpus. So, to avoid the degradation of the performance of the algorithms, comparisons are performed up to four words i.e. Up-to four-gram.

D. Cluster Analysis using K-Means Algorithm

Two classes are proposed to label the type of emails, spam class and ham class. We model the framework for spam detection using classification algorithms assisted by a clustering technique. Instead of asking the users to label the email as spam and ham email, we labelled the emails by using algorithms for two clusters spam cluster and ham cluster. To develop the clustering model, we use K-Means algorithm. Other clustering algorithms like hierarchical clustering and density-based clustering can also be used. Due to a large number of emails in the dataset, classification alone can be a time-consuming step, so in the research work, MLP classification algorithm is assisted by K-Means algorithm for the formation of initial two clusters and then classifying the emails into two classes i.e. Spam class and ham class.

K-Means clustering algorithm is the fastest algorithm that works efficiently on a large dataset. The problem of randomization of the MLP classification algorithm degrades its ability to remove vague information from the dataset. To overcome the disadvantage of randomization of the MLP classification algorithm, initial clusters are provided by the K-Means clustering algorithm.

E. Classification by MLP-NN

Various classification algorithms are present that are used to detect the spam emails such as Decision Tree, Naive Bayes and SVM are used for the detection of spam emails. In our research, we use MLP Neural Network, because of its advantages such as generalisation and fault-tolerance. The main objective of this proposed work is to upgrade the existing machine learning techniques in distinguishing spam emails.

F. Proposed N-K-MLP Algorithm for Gratuitous Mails

Let an Email dataset containing *n* emails to be labelled as spam or ham;

Output: Mails are labelled into two classes as spam or ham.

Stage 1. Perform dataset pre-processing by lexical analysis, removing stop words and stemming.
Stage 2. Calculate the N-Grams for choosing best features for bi-gram, tri-gram and four-gram.
Stage 3. Perform K-Means Clustering for selection of initial clusters and for grouping the emails in two defined clusters viz. spam and ham clusters.
Stage 4. Provide the K-Means results to the MLP model as initial clusters for avoiding randomization for the detection of vague information and classifies the emails in two classes viz. spam and ham.
Stage 5. Compute the output whether an email is a spam or ham.

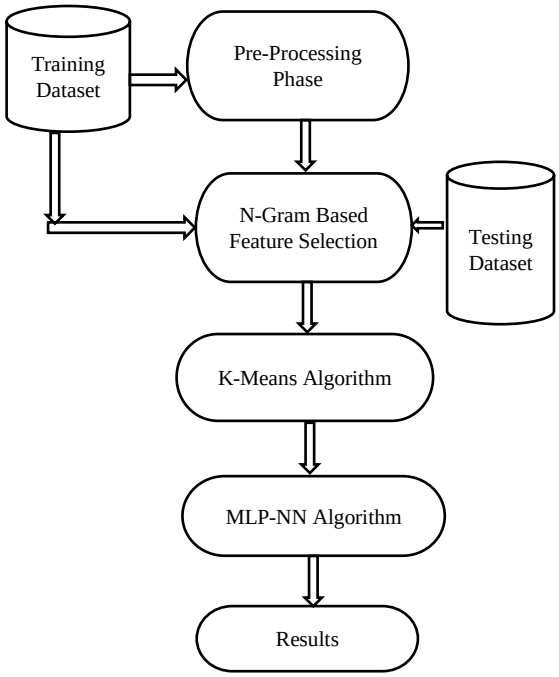


Fig.7. Proposed Technique for the Spam Detection.

V. Results and Discussion

In this section, discussion on the results formulated by the proposed methodology on Enron dataset in contrast to the previous existing methodology is illustrated and

compared. Results ensure that the proposed methodology is more efficient with the highest accuracy and is the most suitable model for distinguishing spam emails. The results are performed on bi-gram, tri-gram and four-gram for various parameters as labelled in Table 1, 2, and Table 3.

Table 1. SVM and MLP Comparison on ENRON Dataset after Pre-Processing.

Percentage Split	Parameters / Algorithms	SVM	MLP
66%	Correctly Classified Instances (Accuracy)	64.66 %	78.09%
66%	In-Correctly Classified Instances	35.34%	21.91%
66%	Sensitivity	0.65	0.781
66%	Specificity	0.489	0.786
66%	Precision	0.722	0.789
66%	F-Measure	0.563	0.783
66%	Root Mean Square Error (RMSE)	0.594	0.386

In Table 1, results are compared for simple MLP and simple SVM classification algorithm after conducting pre-processing of the Enron dataset. The results show better accuracy for MLP algorithm over the SVM algorithm for pre-processed Enron dataset. Fig. 8 illustrates the comparison between both the classification algorithms, where SVM correctly classified 64.66% instances with 35.34% incorrectly labelled instances while MLP performs better with 78.09% accuracy for correctly labelling the instances and 21.91% for incorrectly labelled instances. Fig. 9 shows the comparison for root mean square error, for an algorithm it is desirable to have a low root mean square error. In this case, MLP showed a low error rate of 0.3867 while SVM demonstrated 0.5944 error rate.

Pre-processing of the dataset, eliminates the bogus, missing and incomplete values from the dataset. Secondly, Enron-dataset is a text dataset so it is essential to convert text corpus into words, so as to avoid the performance degradation of the algorithms. The main issue with the MLP classification algorithm is the randomization of the algorithm, that makes the algorithm highly time-consuming and degrades the performance measure of the MLP classification algorithm. In the research work, the main focus is to uplift the performance the MLP algorithm for the detection of email spamming and to remove any kind of vague information by avoiding the randomization of the classification algorithm.

In Table 1, the pre-processing results can clarify that MLP classification algorithm is a better approach for the detection of spam emails with high accuracy of 78.09%. Sensitivity and Specificity rate for the detection of the spam and ham emails for the MLP is high as considered to the SVM technique with 0.781and 0.786 rate respectively.

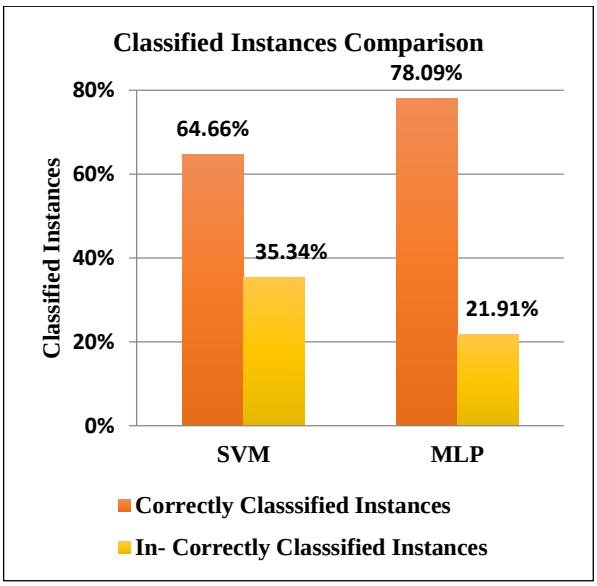


Fig.8. Classified Instances Comparison for SVM and MLP Algorithms.

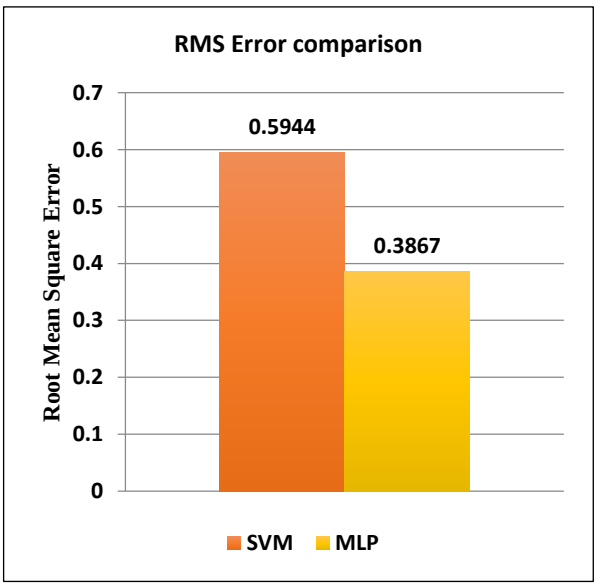


Fig.9. RMS Error Comparison for SVM and MLP Algorithms.

The results are carried on Enron datasets with 1000 emails containing 50% spam and ham rate. Table 1 demonstrates the analysis performed on SVM and MLP classification algorithms. Initially, the pre-processing of the algorithms is performed contributing in the removal of stop words, stemming, and lexical analysis of the dataset, from the refined dataset best features, are extracted depending upon the term frequencies (TF) of the words. The words with frequencies 100 and above are used for analysis. The pre-processed dataset is made available for the classification algorithms viz. SVM and

MLP algorithms, where MLP performed better than SVM classification algorithm. In Table 1, MLP showed 78.09% accuracy with 21.91% incorrectly classified instances. In the next steps, the accuracy of the MLP is enhanced by joining K-Means algorithm.

Table 2. SVM and MLP Comparison after Implementing N-Gram Based Feature Selection Technique.

Percentage Split	Algorithms	SVM ALGORITHM			MLP ALGORITHM		
66%	Parameters	Bi-Gram-SVM	Tri-Gram-SVM	Four-Gram-SVM	Bi-Gram-MLP	Tri-Gram-MLP	Four-Gram-MLP
66%	Correctly Classified Instances (Accuracy)	65.01%	63.95%	63.95%	79.15%	81.62%	79.85%
66%	In-Correctly Classified Instances	34.98%	36.04%	36.04%	20.84%	18.37%	20.14%
66%	Sensitivity	0.622	0.64	0.64	0.792	0.816	0.799
66%	Specificity	0.37	0.360	0.360	0.807	0.838	0.802
66%	Precision	0.387	0.409	0.409	0.806	0.833	0.817
66%	Recall	0.622	0.64	0.64	0.792	0.816	0.799
66%	F-Measure	0.477	0.499	0.499	0.793	0.818	0.8
66%	Root Mean Square Error (RMSE)	0.614	0.600	0.600	0.407	0.392	0.399

The degradation in the performance of the SVM algorithm is due to the failure to accommodate all the data objects lying far from the density function. In Table 2, N-gram based feature selection technique is implemented for selecting the features from the dataset and 100 features are selected for performing the analysis for bi-gram, tri-gram, and four-gram. It was analysed that increasing the value of N (number of words) up to 5 words or above degrades the performance of the algorithms, so the analysis of the N-gram was performed up to four words to avoid the formation of the sentence corpus.

In Table 2, result comparison for N-gram-SVM and N-gram-MLP is performed by assigning the pre-processed and N-gram data to the classification algorithms. Our proposed approach of N-gram-MLP gives better result over N-gram-SVM, but still, the major problem of randomization of the MLP is not solved. To eliminate the bogus data and to boost the performance of the MLP algorithm by discarding the randomization of the algorithm, K-Mean clustering algorithm is joined along with MLP classification algorithm.

In Table 2, we can check the fluctuations in the values of the MLP and SVM algorithm, the values are increasing and decreasing for Bi-gram, Tri-gram and Four-gram. The main reason for the fluctuations in the values of the N-Gram is due to the problem of randomization of the MLP classification algorithm. The main work of the N-Gram based feature selection technique is to select the best features from the text dataset. After selecting the best 100 features from the Enron dataset the improvement in the algorithms is by 1% only. The randomization problem of the MLP algorithm still prevails to boost the performance and accuracy of the algorithm.

In Table 3, the results analysed on the refined MLP using K-Means clustering algorithms is demonstrated. In the initial step, pre-processing of the dataset is carried out that contributes in performing the lexical analysis of the dataset that splits the sentence into words, then removing the stop words is performed in which words like of, the, and that have highest frequencies and is of no use for analysis is removed, lately the stemming of the words is performed and the words that have similar meaning is removed. On the pre-processed dataset, TF-IDF is conducted and the words whose frequency is above 100 is kept for future analysis. The IDF ensure that the word has a greater importance. In the next step, 100 best features are selected by using N-gram based feature selection technique. The pre-processed data is assigned to K-Means clustering algorithm for the formation of initial

clusters for MLP algorithm. K-Means being the fastest clustering algorithm assigns the email data to two defined clusters viz. spam cluster or ham cluster.

Table 3. Comparison of the Proposed N-Gram-K-MLP with the Existing Technique after Implementing K-Means Algorithm.

Percentage Split	Algorithms	N-Gram-K-SVM			Proposed N-Gram-K-MLP		
66%	Parameters	Bi-Gram-K-SVM	Tri-Gram-K-SVM	Four-Gram-K-SVM	Bi-Gram-K-MLP	Tri-Gram-K-MLP	Four-Gram-K-MLP
66%	Correctly Classified Instances (Accuracy)	89.04%	96.81%	97.17%	97.52%	98.23%	99.00%
66%	In-Correctly Classified Instances	10.95%	3.180%	2.82%	2.47%	1.76%	0.90%
66%	Sensitivity	0.89	0.968	0.972	0.975	0.982	0.99
66%	Specificity	0.701	0.840	0.854	0.932	0.911	0.927
66%	Precision	0.905	0.969	0.973	0.976	0.983	0.991
66%	Recall	0.89	0.968	0.972	0.975	0.982	0.99
66%	F-Measure	0.88	0.967	0.971	0.975	0.982	0.99
66%	Root Mean Square Error (RMSE)	0.331	0.178	0.168	0.156	0.127	0.113

The formulated initial clusters are assigned to the MLP algorithm for the classification process, eliminating the randomization of the MLP and upgrading the performance of the algorithm. The classification data is labelled into two classes ham class and spam class. The results of the MLP and SVM classification algorithms are analysed by K-Fold-Cross Validation Model using 10 folds. K-Fold-Cross Validation Model is a model for ensuring the accuracy of the results obtained from the predictive models like classification. Using the validation model the Enron dataset was trained for 9 folds and tested for 1-fold. The results are analysed on 66% splits for training and 34% for testing.

In the research work, our proposed methodology ensures better results over the existing N-Gram-K-SVM technique with 99% accuracy. In the future work, effective measures will be taken to boost the SVM algorithm and the limitation of failing to accommodate all the data objects lying far from the density function will be taken into consideration.

A. Accuracy

The percentage of correctly classified email datasets instances are called as accuracy. Accuracy can be measured as a ratio of correctly classified instances to the total number of instances in the datasets either spam or ham.

$$\text{Accuracy} = \frac{I_{S\to S} + I_{H\to H}}{I_{S\to S} + I_{S\to H} + I_{H\to S} + I_{H\to H}} \quad (5)$$

Where $I_{S\to H}$ = False negative, $I_{H\to H}$ = True negative, $I_{S\to S}$ = True positive, $I_{H\to S}$ = False positive and 'I' represents instances in the dataset.

Table 4. Accuracy Analysis

Algorithms/ Accuracy	N-Gram-K-SVM	N-Gram-K-MLP
Bi Gram	89.05%	97.53%
Tri Gram	96.82%	98.23%
Four Gram	97.17%	99.00%

In Table 4, accuracy comparison is performed between the existing technique i.e. N-Gram-K-SVM and our proposed approach of eliminating the bogus data by avoiding the randomization of the MLP algorithm by joining it with K-Means algorithm viz. N-Gram-K-MLP. Table 4 show that with the increase in the number of the words for the comparison the accuracy of the algorithms is increasing, but the comparison is performed up to four words only to avoid the formation of the sentence corpus. It was analysed that the accuracy of the MLP classification algorithm boosted from 78.09% to 97.53% for Bi-Gram, 98.23% for Tri-Gram, and 99.00% for Four-Gram for classifying the emails as spam and ham email. The results show that the proposed technique of N-Gram with K-Means and MLP (N-Gram-K-MLP) is a better approach for classifying the email dataset.

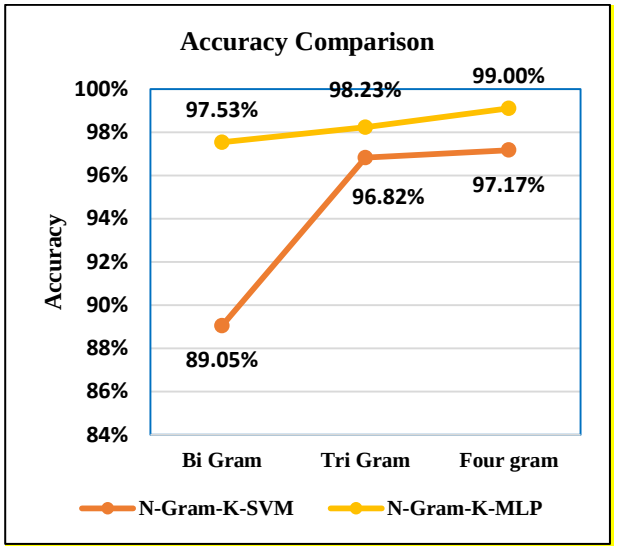


Fig.10. Graphical Representation of Accuracy.

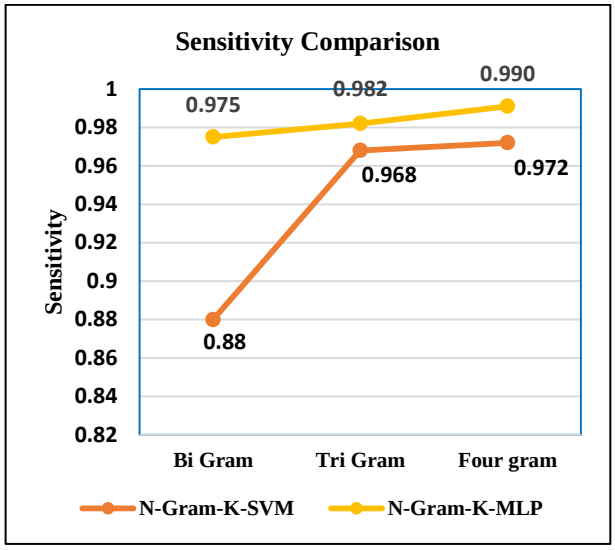


Fig.11. Graphical Representation of Sensitivity.

B. Sensitivity (TP)

The number of emails correctly labelled as spam emails is called as sensitivity. Sensitivity can be calculated by the ratio of the total number of spam emails identified to the total number of spam emails in the dataset.

$$\text{Sensitivity} = \frac{I_{S\to S}}{I_{S\to S} + I_{S\to H}} \quad (6)$$

Where $I_{S\to H}$ = False negative, $I_{S\to S}$ = True Positive and 'I' represents instances in the dataset.

Table 5. Sensitivity Analysis

Algorithms/ Sensitivity	N-Gram-K-SVM	N-Gram-K-MLP
Bi Gram	0.88	0.975
Tri Gram	0.968	0.982
Four Gram	0.972	0.990

Sensitivity also called as True Positive (TP) case or Recall, identifies the number of positive cases. In the research work, sensitivity defines the number of spam emails that are correctly labelled as spam. The above table 5, shows that the proposed approach is highly accurate with 0.975 sensitivity rate for Bi-gram, 0.982 rates for Tri-gram and 0.990 rates for four-gram. It was also analysed that with the increase in the value of N (number of words) the sensitivity rate for the detection of spam emails is also increasing. The existing technique of N-Gram-K-SVM fails to accommodate all the objects lying far from density function, so the existing technique cannot detect all the objects (emails) as spam or ham emails. In the fig. 11, we can clearly analyse that with the increase in the number of words for the N-Gram comparison the performance of the algorithm is continuously increasing with 0.990 sensitivity rate for the detection of spam emails

C. Specificity (TN)

The number of emails correctly labelled to not be a spam mail. Specificity can be calculated by the ratio of the total number of ham emails identified to the total number of ham emails in the dataset.

$$\text{Specificity} = \frac{I_{H\to H}}{I_{H\to S} + I_{H\to H}} \quad (7)$$

Where $I_{H\to H}$= True negative, $I_{H\to S}$ = False Positive and 'I' represents instances in the dataset.

Table 6. Specificity Analysis

Algorithm/ Specificity	N-Gram-K-SVM	N-Gram-K-MLP
Bi Gram	0.701	0.932
Tri Gram	0.84	0.911
Four Gram	0.854	0.927

Specificity also called as True Negative (TN) case, identifies the number of negative cases. In the research work, specificity defines the number of ham emails that are correctly labelled as ham. Sensitivity and specificity are both inversely proportional to each other. It means, if sensitivity is increasing the specificity should decrease and vice versa. In the research work, as shown in Table 6 the specificity is continuously getting lower and showing fluctuation in the results. In Table 5, the sensitivity rate for Bi-gram is 0.975 while specificity rate or true negative rate in Table 6 is decreasing with the value of 0.932 rates. In Table 5, the sensitivity rate for Tri-gram is 0.982 while in table 6 the specificity rate being inversely proportional to the sensitivity is decreasing with the value of 0.911 rate. Similarly, for the comparison of Four-gram, the sensitivity rate and specificity rate is 0.990 and 0.927 respectively. Specificity and sensitivity together detect the number of spam and ham emails by forming a prediction condition for confusion matrix.

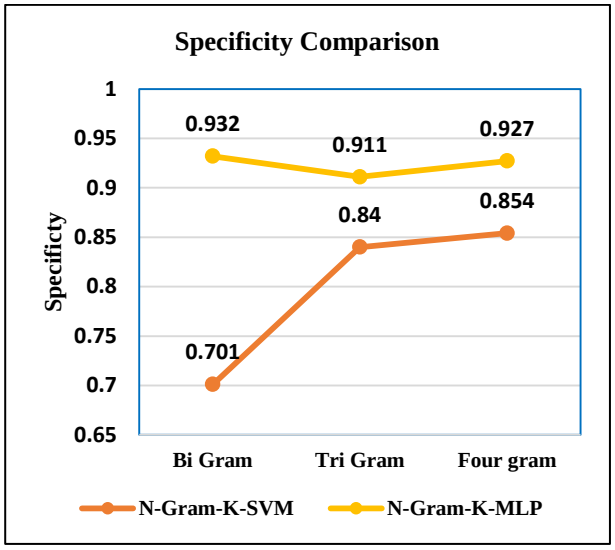


Fig.12. Graphical Representation of Specificity.

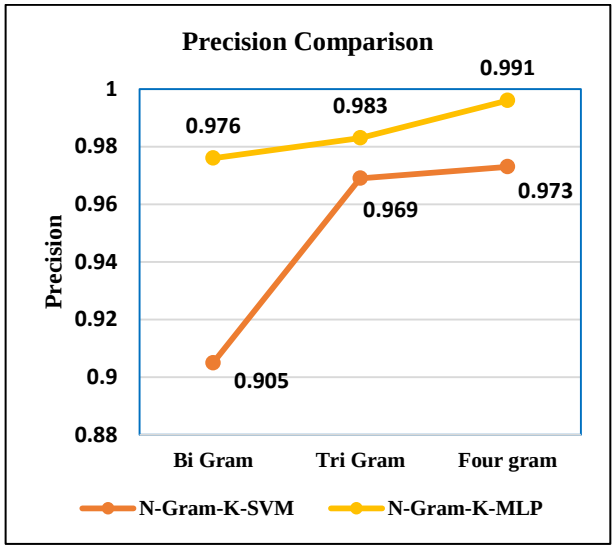


Fig.13. Graphical Representation of Precision.

D. Precision

The prediction ratio of correctly labelled spam emails to the total number of emails correctly labelled as spam.

$$\text{Precision} = \frac{I_{S\to S}}{I_{S\to S} + I_{H\to S}} \quad (9)$$

Where $I_{S\to S}$= True positive, $I_{H\to S}$ = False Positive and 'I' represents instances in the dataset.

Table 7. Precision Analysis

Algorithms/ Precision	N-Gram-K-SVM	N-Gram-K-MLP
Bi Gram	0.905	0.976
Tri Gram	0.969	0.983
Four Gram	0.973	0.991

Precision also called as positive predicted value is the fraction of the relevant documents returned. The formula number 9, defines the precision which defines the ratio of the true positives ie. sensitivity or the number of spam emails detected to the ratio of the false positive (type 1 error) and the true positive. Precision defines the accurate and exact values retrieved. In Table 7, results state that the proposed approach of N-Gram-K-MLP can correctly and accurately retrieve relevant information. Fig 13, clearly shows the continuous increase in the precision rate for correct identification of the emails. The Bi-Gram precision value for the existing technique is 0.905 and for the proposed technique it was found that precision value is 0.976, that means proposed technique can correctly identify the emails as spam and ham emails. For Tri-Gram comparison, the existing technique shows 0.969 precision rate and for proposed technique 0.983 precision rate is detected. Similarly, for Four-Gram the proposed technique show 0.991 precision rate which is higher than the existing technique that shows 0.973 precision rate.

E. Root Mean Square Error

The difference between the values predicted by an algorithm 'y' and the values actually observed from the environment 'y_i' is termed as Root mean square error.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y-yi)^2}{n}} \quad (10)$$

Where 'y' is the predicted value, 'y_i' is the observed values, and 'n' is the number of observations.

Table 8. RMS Error Analysis.

Algorithms/ RMS	N-Gram-K-SVM	N-Gram-K-MLP
Bi Gram	0.331	0.1564
Tri Gram	0.1783	0.1273
Four Gram	0.1681	0.1134

Root mean square error is an efficient and common metrics used to measure the exactness of the continuous variables. Root mean square error is a difference between the actual values predicted and observed values, used for measuring the average magnitude of the errors. A Root Means Square Error (RMSE), has a higher weight for large error values because the errors are firstly squared than later on their average is performed. RMSE penalise the large errors, so it is one of the best kind of error to be predicted. The research work shows that the proposed approach has a lower error rate that the existing technique, hence making the proposed technique as a better model for the classification of spam emails. The existing technique of N-Gram-K-SVM demonstrates 0.331 RMSE while our proposed technique shows a low error rate of 0.1564. In Table 8, the results clearly show that with the increase in the value of N (number of words) the error rate is decreasing continuously making the proposed

approach an efficient classification model.

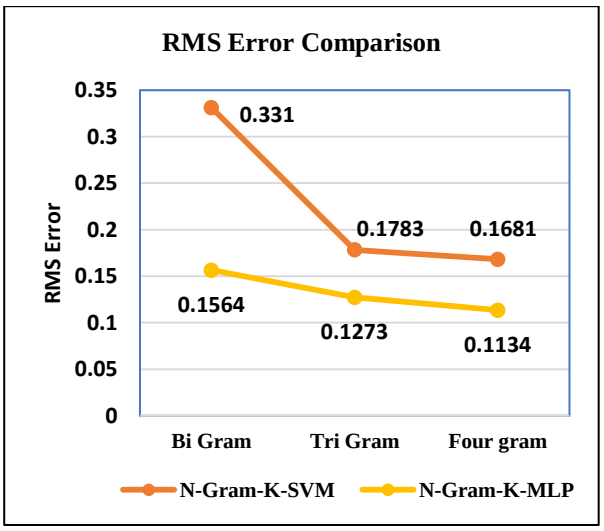


Fig.14. Graphical Representation of RMS Error.

VI. Conclusion

In this paper, efficient and effective analysis of spam email filtration is conducted using joined approach for classification and clustering along with N-gram. Result comparison is performed on bi-gram, tri-gram, and four-gram which clearly illustrate that K-MLP based approach along with N-Gram feature selection technique produces more meaningful and informative clusters for classification. Various studies conducted so far shows that K-Means algorithm is the fastest unsupervised approach that can efficiently work on the large dataset without overlapping and is resistant to noise and outliers. Considering the rapid increase of spammers and spam emails, it is essential to use defensive mechanisms. The problem of randomization of MLP neural network lead to degradation of the performance of the algorithm for the removal of vague information but when MLP is refined using K-Means algorithm it helps the neural network for selecting initial clusters that lead to fast computation for model building of the algorithm and boosted the performance too. The results of simple MLP and SVM is initially carried out which shows the accuracy of 78.09% and 64.66% respectively, though MLP shows higher accuracy than SVM because of the limitation of the SVM algorithm to choose information point because of kernel trick it fails to accommodate all the information point lying far from density function, but still such low accuracy is not suitable for an efficient spam detection model. When we implemented our proposed model of N-K-MLP (K-Means-MLP with N-Gram) it demonstrated higher performance than existing N-K-SVM (K-Means-SVM with N-Gram) along with low error rate and the performance of the MLP was boosted to 97.53% for Bi-Gram, 98.23% for Tri-Gram and 99.00% for 4-Gram. N-Gram helped in choosing the best features from the large dataset. Our proposed model gives better results over simple MLP, simple SVM, and N-gram based K-SVM. In the research work, we performed N-gram analysis up to four words to avoid the formation of sentence corpus, as increasing the value of N (number of words) slower the performance of the algorithm.

Acknowledgment

The author expresses its humble thanks to CT Group of Engineering, Management, and Technology for their motivational participation and encouragement in the research field. The author also presents its gratitude towards computer science research group for the support. The author is thankful to the mentor for guidance throughout the research work.

References

[1] B. Yu and Z. Xu, "A comparative study for content-based dynamic spam classification using four machine learning algorithms", *Knowledge Based System-Elsevier, vol. 21, pp. 355–362, 2008*.

[2] T. A. Almeida and A. Yamakami, "Content-Based Spam Filtering", *in International Joint Conference on Neural Networks (IJCNN) - IEEE, pp. 1-7, 2010.*

[3] L. Firte, C. Lemnaru, and R. Potolea, "Spam Detection Filter using KNN Algorithm and Resampling", *in 6th International Conference on Intelligent Computer Communication and Processing -IEEE*, pp.27-33, 2010.

[4] D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques", *in 2011 International Conference on Process Automation, Control and Computing - IEEE*, pp. 1–7, 2011.

[5] R. Shams and R. E. Mercer, "Classifying spam emails using text and readability features", *in International Conference on Data Mining (ICDM) - IEEE*, pp. 657–666, 2013.

[6] M. Rathi and V. Pareek, "Spam Email Detection through Data Mining - A Comparative Performance Analysis", *International Journal of Modern Education and Computer Science (IJMECS)*, vol. 12, pp. 31-39, 2013.

[7] A. Harisinghaney, A. Dixit, S. Gupta, and Anuja Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm", *in International Conference on Reliability, Optimization and Information Technology (ICROIT)-IEEE*, pp.153-155, 2014.

[8] S. P. Teli and S. K. Biradar, "Effective Email Classification for Spam and Non- spam", *International Journal of Advanced Research in Computer and Software Engineering*, vol. 4, 2014.

[9] Alsmadi and I. Alhami, "Clustering and classification of email contents", *Journal of King Saud University - Computer and Information Science -Elsevier*, vol. 27, no. 1, pp. 46–57, 2015.

[10] A. S. Aski and N. K. Sourati, "Proposed efficient algorithm to filter spam using machine learning techniques", *Pacific Science Review- A Natural Science Engineering- Elsevier.*, vol. 18, no. 2, pp. 145–149, 2016.

[11] M.Prilepok and P. Berek, "Spam Detection Using Data Compression And Signatures And Signatures," *Cybernetics and Systems: An International Journal*, vol. 44, pp. 533–549, 2014.

[12] G. Kaur, R. K. Gurm, "A Survey on Classification Techniques in Internet Environment", *International Journal of Advance Research in Computer and Communication Engineering (IJARCCE)*, vol. 5, no. 3, pp. 589–593, March 2016.

[13] Rekha and S. Negi, "A Review on Different Spam Detection Approaches", *International Journal of Engineering Trends and Technology (IJETT)*, vol.11, no.6, 2014

[14] Z. Elberrichi and B. Aljohar, "N-grams in Texts Categorization," S*cientific Journal of King Faisal University (Basic and Applied Sciences)*, vol. 8, no. 2, pp. 25–39, 2007.

[15] D. Jurafsky and J. H. Martin, "N-Gram," *Speech and Language Processing*, 2014.

[16] J. Clark, I. Koprinska and J.Poon, "A Neural Network-Based Approach to automated email classification", *in WIC International Conference on Web Intelligence – IEEE*, 2003.

[17] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. J. Rajabi, "Advantage and drawback of support vector machine functionality," in *1st International Conference on Computer Communication and Control Technology - IEEE*, pp. 63–65, 2014.

[18] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," *Technical Report -Microsoft Research*, 2010.

[19] M. Iqbal, M. M. Abid, M. Ahmad, and F. Khurshid,"Study on the Effectiveness of Spam Detection Technologies", *International Journal of Information Technology and Computer Science (IJITCS)*, Vol.8, No.1, pp.11-21, 2016.

[20] R. Xu. and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Transaction on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[21] M. S. Chen, J. Han, and P. S. Yu, "Data Mining: An overview from a database perspective," *IEEE Transaction on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.

[22] J. Dermoudy, B. Kang, D. Bhattacharyya, et. al. "Process of Extracting Uncover Patterns from Data: A Review," *International Journal of Database Theory and Application (IJDTA)*, Vol. 2, No. 2, June 2009.

[23] D. Guan, W. Yuan, Y. Lee, A. Gavrilov, and S. Lee, "Combining Multi-Layer Perceptron and K-means for Data Clustering with Background Knowledge," *Springer*, pp. 1220–1226, 2007.

[24] P. Verma and D. Kumar, "Association Rule Mining Algorithm's Variant Analysis," *International Journal of Computer Appl*ication *(IJCA)*, vol. 78, no. 14, pp. 26–34, 2013.

[25] R. Xu. and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Transaction on. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[26] M. S. Chen, J. Han, and P. S. Yu, "Data Mining: An overview from a database perspective," *IEEE Transaction on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.

[27] A. Silberschatz, M. Stonebraker and J.D. Ullman, "Database Research: Achievements and Opportunities into the 21st Century," *Report NSF workshop Future of Database Systems Research*, May 1995.

[28] G. Piatetsky Shapiro and W.J. Frawley, "Knowledge Discovery in Databases", *AAAI/MIT Press*, 1991.

[29] A. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[30] C. Bishop, Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995.

[31] A. Baraldi and E. Alpaydin, "Constructive Feedforward ART clustering networks—Part I and II," *IEEE Transaction Neural Network.*, vol. 13, no. 3, pp. 645–677, May 2002.

[32] V. Cherkassky and F. Mulier, Learning from Data: Concepts, Theory and Methods. New York: Wiley, 1998.

[33] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2ND ED. New York: Wiley, 2001.

[34] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[35] M. S. B. PhridviRaj and C. V. GuruRao, "Data Mining – The Past, Present and Future – A Typical Survey on Data Streams," *Procedia Technology*, vol. 12, pp. 255–263, 2014.

[36] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazines*, vol. 17, no. 3, p. 37, 1996.

[37] B. Everitt, S. Landau, and M. Leese, Cluster Analysis. London: Arnold, 2001.

[38] A. Jain, A. Rajavat, and R. Bhartiya, "Design, analysis and implementation of modified K-mean algorithm for the large dataset to increase scalability and efficiency," *In - 4th International Conference on Computer Intelligence and Communication Networks (CICN)*, pp. 627–631, 2012.

[39] P. Chauhan and M. Shukla, "A Review on Outlier Detection Techniques on Data Stream by Using Different Approaches of K-Means Algorithm," *In - International Conference on Advances in Computer Engineering and Applications (ICACEA)*, pp. 580–585, 2015.

[40] S. Firdaus and A. Uddin, "A Survey on Clustering Algorithms and Complexity Analysis," *International Journal of Computer Science Issues (IJCSI)*, vol. 12, no. 2, pp. 62–85, 2015.

[41] D. Sisodia, "Clustering Techniques: A Brief Survey of Different Clustering Algorithms," *International Journal on latest trends and Engineering Technology(IJLTET)*, vol. 1, no. 3, pp. 82–87, 2012.

[42] K. N. Ahmed and T. A. Razak, "An Overview of Various Improvements 8of DBSCAN Algorithm in Clustering Spatial Databases," *International Journal of Advance Research in Computer and Communication Engineering (IJARCCE)*, vol. 5, no. 2, pp. 360–363, 2016.

[43] A. Joshi, "A Review: Comparative Study of Various Clustering Techniques in Data Mining," *International Journal of Advance Research in Computer Science and Software Engineering (IJARCSSE)*, vol. 3, no. 3, pp. 55–57, 2013.

[44] A. Naik, "Density Based Clustering Algorithm," 06-Dec-2010.[Online].Available:https://sites.google.com/site/dataclusteringalgorithms/density-based-clustering-algorithm. [Accessed: 15-Jan-2017].

[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *SIGKDD Exploration Newsletter.*, vol. 11, no. 1, pp. 1-10, 2009.

[46] R. Ng and J. Han," Efficient and Effective Clustering Method for Spatial Data Mining," *In - 20th VLDB Conference*, pp. 144-155, Santiago, Chile,1994.

[47] Cios, K. J., W. Pedrycz, et al., *Data Mining Methods for Knowledge Discovery*, vol. 458, Springer Science &

Business Media, 2012.
[48] S. Dixit, and N. Gwal, "An Implementation of Data Pre-Processing for Small Dataset," *International Journal of Computer Application (IJCA)*, vol. 10, no. 6, pp. 28-3, Oct. 2014.
[49] S. Singhal and M. Jena, "A Study on WEKA Tool for Data Pre-processing, Classification and Clustering," *International Journal of Innovative Technology and Exploration Engineering*, vol. 2, no. 6, pp. 250–253, May 2013.
[50] O. Y. Alshamesti, and I. M. Romi, "Optimal Clustering Algorithms for Data Mining*" Int. Journal of Info. Eng. and Electron. Bus. (IJIEEB)*, vol. 5, no. 2, pp. 22-27, Aug 2013. "DOI: 10.5815/ijieeb.2013.02.04 "
[51] N. Lekhi, M. Mahajan "Outlier Reduction using Hybrid Approach in Data Mining," *International Journal of Modern Education and Computer Science (IJMECS)*, vol. 7, no. 5, pp. 43–49, May 2015.
[52] C. L. P. Chen and C.Y. Zhang, "Data- Intensive Applications, Challenges, Techniques and Technologies: A survey on Big Data." *ELSEVIER- Information Science*, pp. 314-347, Aug. 2014.
[53] E. Rahm, and H. H. Do, "Data Cleaning: Problems and current approaches," *IEEE- Data Engineering Bulletin*, vol. 23, no. 4, pp. 3-13, Dec 2000.
[54] H Kaur, P. Verma, "Survey on E-Mail Spam Detection Using Supervised Approach with Feature Selection," *International Journal of Engineering Sciences and Research Technology (IJESRT), vol. 6, no. 4, pp. 120-128, April 2017.*

Authors' Profiles

Harjot Kaur was born in Jalandhar, Punjab, India in 1992. She received the B.Tech degree in Computer Science and Engineering from C.T. Group of Institution, Jalandhar, India, in 2014. She is currently a student of M.Tech in Computer Science and Engineering from C.T. Group of Institution, Jalandhar, India. The M.Tech degree will be completed in 2017. Her main areas of research interests are Data Mining and Data Warehousing.

Prince Verma, he received the B.Tech degree in Computer Science from MIMIT, Malout (Pb), India in 2008 and M.Tech degree in Computer Science in 2013 from DAVIET, Jalandhar (Pb), India. Currently, he is Assistant Professor in Computer Science Department of CTIEMT, Jalandhar (Pb), India. His research focuses on Data Mining and Algorithm optimisation technique.

Manuscript received May 09, 2017; revised May 17, 2017; accepted June 05, 2017.

How to cite this paper: Harjot Kaur, Er. Prince Verma,"K-MLP Based Classifier for Discernment of Gratuitous Mails using N-Gram Filtration", International Journal of Computer Network and Information Security(IJCNIS), Vol.9, No.7, pp.45-58, 2017.DOI: 10.5815/ijcnis.2017.07.06