International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 13-19
Published Online April 2018 in IJEAST (http://www.ijeast.com)

FEATURE SELECTION BASED EFFICIENT MACHINE LEARNING TECHNIQUE FOR EMAIL SPAM PREDICITION

Jagdeep Kaur, Priyanka
Computer Science and Engineering
Swami Vivekanand Institute of Engineering and Technology

Abstract- **Electronic mail (E-mail) has become the lifeline of the majority of modern business as well as a common vehicle for interpersonal communication between connected people. E-mail is so popular, since it is simple, cost effective and supports nearly instantaneous delivery. With such an increase in use of E-mail as a means of communication, the volume of unwanted email messages (mostly spam E-mail) that is received, annually, has also grown significantly. Spam E-mails have begun to gradually undermine the integrity of E-mail and degrade online experience. This paper propose a technique to detect spam emails using improved MLP with N-gram feature selection. Results of the proposed technique is analyzed and compared with the existing technique on the basis of Accuracy, Recall, Fmeasure, Root Mean Square Error and Precision.**

Keywords **– Email, Spam, MLP, N-gram feature selection, k-means clustering.**

I. INTRODUCTION

Email is the most effective and speediest method of correspondence to trade data over the web. Because of the expansion in the quantity of record holders over the different social locales, there is an enormous increment in the rate of spreading of spam messages. Regardless of having different instruments accessible still, there are many hotspots for the spam to start. Absence of guard instrument to keep the spreading of spam can cause serious financial misfortune, loss of data transmission for taking care of spam messages, memory use and can make individual and money related dangers the data holders. Spam can be comprehended as 'an undesirable ill-conceived, garbage messages got by the authentic clients from unauthenticated sources'. To deal with spam messages spam filtration strategy is taken after which obstructs the spam mail from going into the mail inbox, yet the significant issue with spam filtration is that a substantial email can be distinguished as spam or a spam email can be missed. Spam can be sifted by a non-machine learning and machine learning methods.

The 'E-mail Spam' doesn't yet has a universally accepted formal definition, but it can be broadly described as unwanted E-mail message(s) or unsolicited commercial E-mail(s). Of late, the volume of certain types of E-mails has significantly increased, which, strictly speaking, may not be necessarily considered to be commercial in nature, but are sent in bulk without consent (expressed or implied) of recipients. Thus, they too are considered to be spam E-mails

II. BACKGROUND

B. Yu et al. played out a relative examination on content-based spam grouping utilizing four distinctive machine learning calculations. This paper grouped spam messages utilizing four distinctive machine learning calculations viz. Naıve Bayesian, Neural Network, Support Vector Machine and Relevance Vector Machine. The investigation was performed on the diverse preparing dataset and highlight determination. Investigation comes about showed that NN calculation is no sufficient calculation to be utilized as an instrument for spam dismissal. SVM and RVM machine learning calculations are preferable calculations over NB classifier. Rather than moderate learning, RVM is still preferable calculation over SVM for spam characterization with less execution time and less pertinence vectors. [1]

A. Almeida et al. examined a similar examination utilizing content-based separating for spam. This paper talked about seven distinctive changed variants of Naïve Bayes Classifier and contrasted those outcomes and Linear Support Vector Machine on six diverse open and huge datasets. The outcomes exhibited that SVM, Boolean NB and Basic NB are the best calculations for spam discovery. In any case, SVM executed the precision rate higher than 90% for all the datasets used. [2]

Loredana Firte et al. demonstrated a relative investigation on spam identification channel utilizing KNN Algorithm and Resampling approach. This paper makes utilization of the K-NN calculation for grouping of spam messages on the predefined dataset utilizing highlight's chosen from the substance and messages properties. Resampling of the datasets to suitable set and positive appropriation was done to make the calculation effective for highlight choice. [3]

Aurangzeb Khan et al. spoken to an audit of the hypothesis and methods of content mining and archive

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 13-19
Published Online April 2018 in IJEAST (http://www.ijeast.com)

arrangement and focusing on the current writing. This paper give a knowledge of machine learning procedures and reports portrayal strategies. They reasoned that Gain and chi square technique for highlight determination is ideal and for the most part utilized among others. They exhibited different characterization strategies for test mining and report choice and furthermore clarify some half and half techniques for that which might be the blend of at least two from the current strategies. The creator reasoned that the Naïve Bayes perform well for spam sifting and email classification. [4]

D. K. Renuka et al. talked about a near investigation of spam characterization in view of regulated getting the hang of utilizing a few machine learning strategies. In this investigation, the correlation was finished utilizing three distinctive machine learning order calculations viz. Guileless Bayes, J48 and Multilayer perceptron (MLP) classifier. Results showed high exactness for MLP however high time utilization. While Naïve Bayes exactness was low than MLP however was sufficiently quick in execution and learning. The precision of Naïve Bayes was upgraded utilizing FBL include determination and utilized sifted Bayesian Learning with Naïve Bayes. The adjusted Naïve Bayes demonstrated the precision of 91%. [5]

RasimM. Alguliev et al. explained on this paper, the crisis of clustering of junk mail messages assortment is formalized. The criterion function is a max of similarity between messages in type of clusters, which is defined through ok-nearest neighbor algorithm. They use genetic algorithm including penalty perform for solving clustering trouble. And then classification utilising ok-nearest neighbor algorithm was once applied to monitor spam emails in each of the cluster. After that Multi file summarization process is utilized for competencies extraction from clusters. The know-how which was once retrieved from every clusters and thematic dependence of junk mail emails from their starting place can also be valuable in accumulating data about social networks of spammers if any. [6]

Aman Kumar Sharma et al. carried out work to seek out the accuracy of classification of 4 algorithms for e mail to notice unsolicited mail or now not. They makes use of WEKA, the data mining software for their experiment. They use ID3, J48, simple CART and ADTree, to compare the accuracy. They acquire the dataset of 4601 emails in whole with fifty eight attributes. After their scan they concluded that J48 is the first-class algorithm among the others which offers the accuracy of 92.76%. Easy CART has very near accuracy of J48 i.e. 92.63%. [7]

Rushdi Shams et al. carried out a comparative analysis of the classification of junk mail emails by utilising text and readability aspects. This paper proposed an effective unsolicited mail classification system along with feature decision making use of the content of emails and readability. This paper used 4 datasets reminiscent of CSDMC2010, spam murderer, Ling spam, and Enron-spam. Facets are labeled into three classes i.e. Natural features, test elements and readability points. The proposed process is equipped to categorise emails of any language when you consider that the aspects are stored impartial of the languages. This paper used five classification centered algorithms for spam detection viz. Random wooded area (RF), Bagging, AdaBoost, aid Vector machine (SVM) and Naïve Bayes (NB). Results comparison among one of a kind classifiers envisioned Bagging algorithm to be the excellent for unsolicited mail detection. [8]

III. PROPOSED TECHNIQUE

Email is primary the accepted adjustment of advice today over the internet and emails can be spam or ham. Spam emails are beatific to the recipients in aggregate and are exceptionable to receiver. These types of spams are actual austere and actionable to the recipients. As the internet users are numerously accretion day by day, it is accordingly important to administer the emails and adopting problems of abusage amid bodies and organizations. The aggregate of exceptionable letters to almsman categorized as SPAM is an archetype of email misuse. Email is the accepted anatomy of spamming on the internet. A above call is to assure the user from the spam mails. Various allocation algorithms are present that are acclimated to ascertain the spam mails.

This spam filtration botheration is not new and abounding allocation algorithms such as Decision tree, Naive Bayes, SVM, Neural networks had been acclimated in altered types of training datasets and they had been accustomed acceptable allocation after-effects as able-bodied as bigger efficiencies.

The existing methods have some limitations of having less accuracy and precision. In the base paper, inbuilt algorithms are used in order to do the classification; From those algorithms MLP is having highest accuracy rate. MLP algorithm make initial clusters based on the randomized approach which need to eliminate vague information. Therefore, in order to remove this clustering problem, the initial clusters are created using nearest neighbor approach. The nearest neighbor is find using distance formula. The main purpose of this proposed work is to promote the presented machine learning methods in distinctive spam emails. The objectives of the proposed technique are:

- To study various spam detection algorithms for emails.
- To propose an approach for email spam detection using improved MLP with N-gram feature selection.
- To compare and analyse the results of proposed approach with the existing on the basis of parameters viz. Accuracy, Recall, Fmeasure, Root Mean Square Error and Precision.

The methodology for proposed technique is as follows:

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 13-19
Published Online April 2018 in IJEAST (http://www.ijeast.com)

The suggested technique comprises from claiming Different steps: (1) Dataset pre-processing (2) characteristic determination utilizing N-Gram (3) group examination Toward K-Means (4) arrangement by MLP. Correlation of the suggested method might have been conveyed crazy for the existing approach which utilization MLP algorithm to order of spam messages. Those effects were directed with respect to Enron dataset by lessening the Characteristics to better Investigation.

Let an Email dataset absolute 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 allotment best appearance for bi-gram, tri-gram and four-gram.

Stage 3: Perform K-Means Clustering for alternative of antecedent clusters and for alignment the emails in two authentic clusters viz. spam and ham clusters.

Stage 4: Provide the K-Means after-effects to the MLP archetypal as antecedent clusters for alienated randomization for the apprehension of ambiguous advice and classifies the emails in two classes viz. spam and ham.

Stage 5: Compute the achievement whether an email is a spam or ham.

IV. EXPERIMENTAL RESULTS

The simulation has been done in Java Net Beans. NetBeans is an open-source project dedicated to providing rock solid software development products (the NetBeans IDE and the NetBeans Platform) that address the needs of developers, users and the businesses who rely on NetBeans as a basis for their products; particularly, to enable them to develop these products quickly, efficiently and easily by leveraging the strengths of the Java platform and other relevant industry standards.

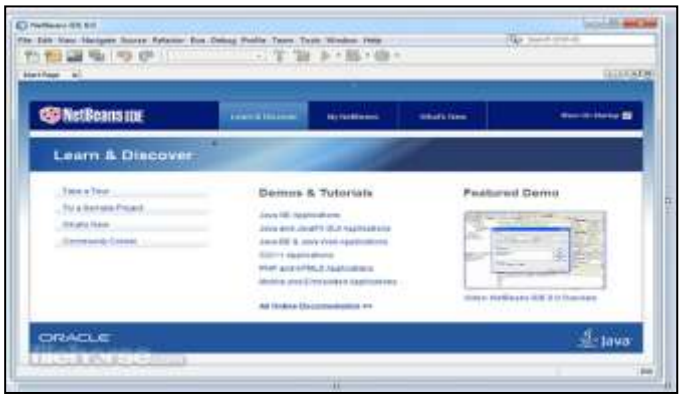

Figure 4.1 NetBeans IDE

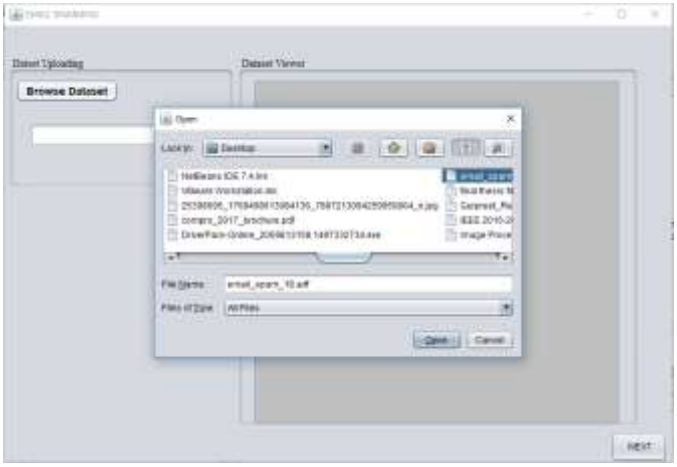

Figure 4.2 Dataset Selection

This is the first window we come across,here we select the files that our dataset have.that are preloaded when we create database at the backend.We have created emails. arff file .Here's it contains our emails dataset that we browse through,we upload them as such.here we upload from the main server thing. This interface helps us to choose the desired data set from any location and upload that data set. The data set which we upload comes under the files list The purpose of browsing is to select the dataset from any location within the computer so that it will be used for the further processing of the dataFrom the files we select the dataset and then whole details and data is loaded under the dataset. After the loading of the dataset the next step is to perform filtration on the text data in order to convert string into words.

Figure 4.3 Visualization of dataset in weka Tools editor

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 13-19
Published Online April 2018 in IJEAST (http://www.ijeast.com)

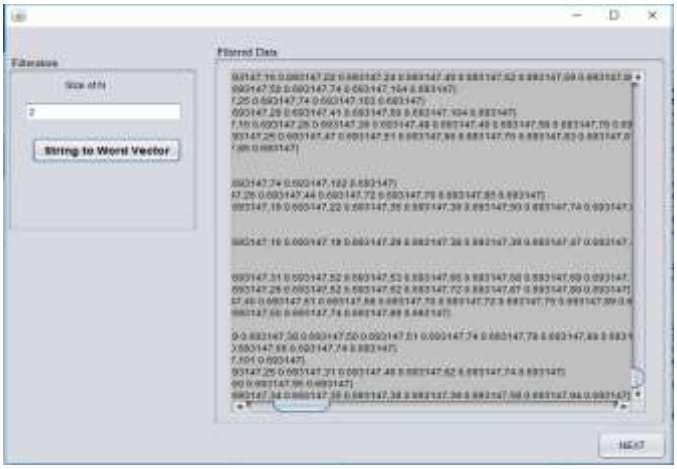

Figure 4.4 Showing the results of String to word vector filter

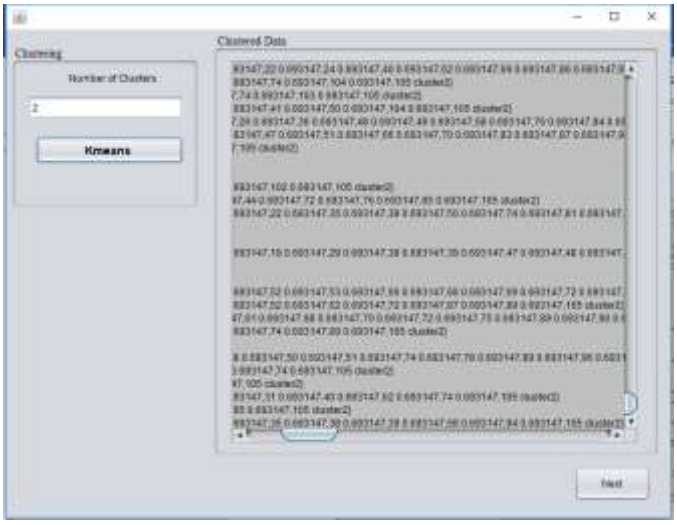

Figure 4.5 Showing the results of Kmeans clustering

The figure above the shows the results of Kmeans algorithm. The data is clustered based on the similarity between the features using minimum Euclidean distance. The existing Multi-layer perceptron algorithm makes random clusters; due to this the accuracy is reduced and data is not efficiently classified. Therefore in the proposed MLP algorithm, data is first clustered with the help of Kmeans clustering algorithm, then this clustered data is classified using MLP algorithm that gives more efficient results.

Figure 4.6 Showing the results of MLP classification algorithm

The figure above shows the classification results of Multilayer Perceptron algorithm. The results show the accuracy of 76.32% i.e. 216 instances are correctly classified out of 283 instances. The multilayer perceptron (MLP) is a feed-forward, supervised learning network with up to two hidden layers. The MLP network is a function of one or more predictors (also called inputs or independent variables) that minimizes the prediction error of one or more target variables (also called outputs). Predictors and targets can be a mix of categorical and scale variable. The kappa statistics for naïve bayes algorithm is 0.5166 Class details parameters are also shown like precision which is 0.768, recall 0.763, F Measure 0.765, TP Rate 0.763 and FP rate 0.239.

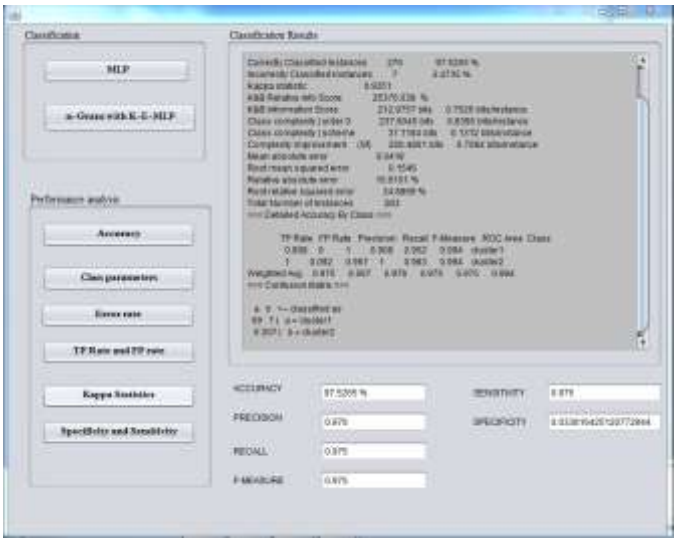

Figure 4.7 Showing the results of n- gram based improved MLP classification algorithm

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 13-19
Published Online April 2018 in IJEAST (http://www.ijeast.com)

Table 4.1: Accuracy comparison of MLP and Proposed MLP on Enron Dataset

Algorithm	Accuracy
MLP	0.5166
N-Gram with K-E-MLP	0.9351

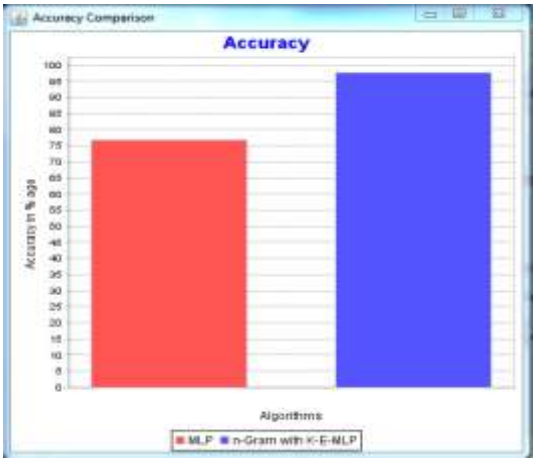

Figure 4.8 Showing the accuracy comparison of MLP and N gram based Proposed MLP

Table 4.2: Class Parameters of MLP and Proposed MLP on Enron Dataset

Class Parameters	MLP	N-Gram with K-E-MLP
Precision	0.768	0.976
Recall	0.763	0.975
F-Measure	0.765	0.975

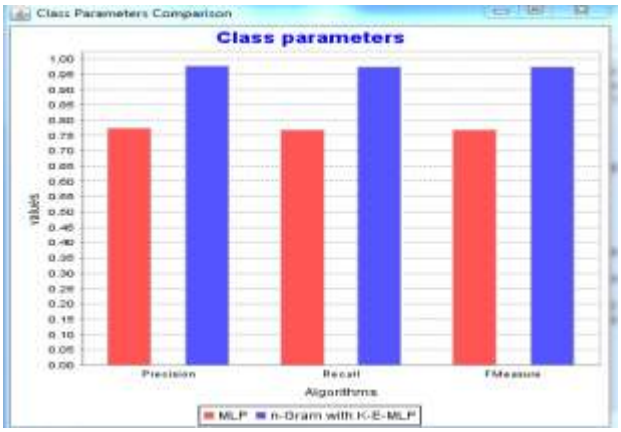

Figure 4.9 Showing the class Parameters comparison of MLP and N gram based Proposed MLP

Table 4.3: Kappa Statistic comparison of MLP and Proposed MLP on Enron Dataset

Algorithm	Accuracy
MLP	76.33
N-Gram with K-E-MLP	97.53

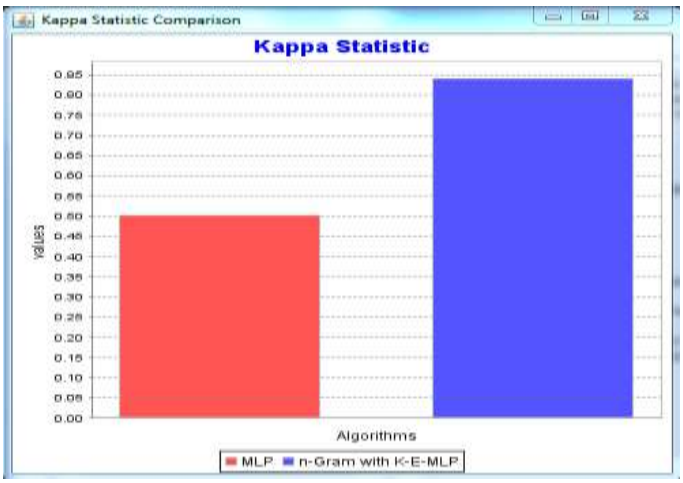

Figure 4.10 Showing the Kappa Statistic comparison of MLP and N gram based Proposed MLP

Table 4.4: Error rate comparison of MLP and Proposed MLP on Enron Dataset

Parameters	MLP	N-Gram with K-E-MLP
Mean absolute error	0.2438	0.0409
Root mean square error	0.4384	0.1546

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 13-19
Published Online April 2018 in IJEAST (http://www.ijeast.com)

Figure 4.11 Showing the error rate comparison of MLP and N gram based Proposed MLP

Table 4.5: TP rate and FP rate comparison of MLP and Proposed MLP on Enron Dataset

Parameters	MLP	N-Gram with K-E-MLP
TP rate	0.763	0.975
FP rate	0.239	0.067

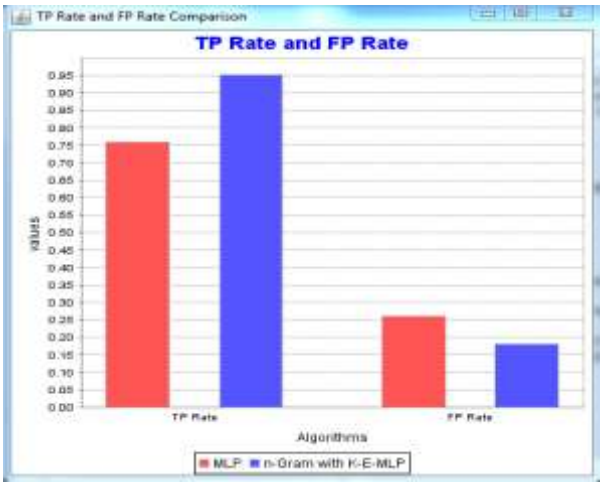

Figure 4.12 Showing the TP Rate and FP Rate comparison of MLP and N gram based Proposed MLP

V. CONCLUSION

In the research work, 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 emails collected from Enron dataset illustrate that n gram based K-Enhanced MLP approach produce 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 large dataset without overlapping and is resistant to noise and outliers. Considering the rapid increase of spammers and spam mails, 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 is initially carried out which shows the accuracy of 76.33%. The proposed n gram based K-Enhanced MLP demonstrated higher performance than existing MLP along with low error rate and the performance of the MLP was boosted to 97.53% for Bi-Gram analysis. N-Gram helped in choosing the best features from the large dataset. The proposed model gives better results over MLP. In future work, results analysis comparison for five-gram and above will be considered and better algorithms will be selected that will enhance the performance of the proposed technique for five-gram and above. Also, some other optimization techniques i.e. genetic algorithm, fuzzy logic, Ant Colony Optimization (ACO) or Artificial Bee Colony (ABC) can be applied to the MLP-NN for the selection of initial clusters. Also, other feature selection techniques like information gain, co-relation feature selection can be implemented for selecting the best features for numerical data, text data and image corpus.

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 of 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] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee and Khairullah khan "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances in Information Technology vol. 1, No.1, February 2010

[5] 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.

[6] Rasim, MA and Ramiz, MA and Saadat, AN," Classification of Textual E-mail spam using Data Mining Techniques", Journal of Applied Computational Intelligence and Soft Computing, 2011.

[7] Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering (IJCSE) , Vol. 3, No. 5, May 2011

[8] 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.

[9] 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.

[10]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.

[11] 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.

[12] Anjali Sharma, Manisha, Dr.Manisha and Dr.Rekha Jain, "A survey on spam detection techniques", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 12, December 2014.

[13] Prachi Goyal Juneja, R. K. Pa teriya, "A Survey on Email Spam Types and Spam Filtering Techniques", International Journal of Engineering Research & Technology, Vol. 3 Issue 3, 2014, pp. 2309-2314.

[14] Rekha, Sandeep Negi, "A Review on Different Spam Detection Approaches", International Journal of Engineering Trends and Technology, Volume 11 Number 6, 2014, pp. 315-318.

[15] Sarju S, Riju Thomas, Emilin Shyni C, "Spam Email Detection using Structural Features", International Journal of Computer Applications, Volume 89 – No.3, 2014, pp. 38-41.

[16] Savita Teli, Santoshkumar Biradar, "Effective Spam Detection Method for Email", IOSR Journal of Computer Science, 2014, pp. 68-72.

[17] 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.

[18] Masurah Mohamad and Ali Selamat, "An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification", International Conference on Computer, Communications, and Control Technology (I4CT), IEEE, 2015.

[19] Priyanka Sao, Pro. Kare Prashanthi, "E-mail Spam Classification Using Naïve Bayesian Classifier", International Journal of Advanced Research in Computer Engineering & Technology, Volume 4 Issue 6, 2015, pp. 2792-2796.

[20] Wazir Zada Khan, Muhammad Khurram Khan, Fahad Bin Muhaya, Muhammad Y Aalsalem, HanChieh Chao, "A Comprehensive Study of Email Spam Botnet Detection", IEEE, 2015.