I.J. Modern Education and Computer Science, **2016, 4, 57-63**
Published Online April 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.04.07

E-Mail Spam Detection Using SVM and RBF

Reena Sharma
Chandigarh University/Computer Science, Mohali, 160055, India
Email: sharmareena551@gmail.com

Gurjot Kaur
Chandigarh University/Computer Science, Mohali, 160055, India
Email: randhawa789@gmail.com

Abstract—In today's life internet is an important part. We spend most of our time on internet. One of the important features of internet is communication. Email is a mode of communication which is used for the personal and business purpose. Spam emails are the emails recipient does not wish to take delivery of; it is also called unwanted bulk email. Emails are used each day by number of user to converse around the world. At present large volumes of spam emails are reasoning serious trouble for Internet user and Internet service. Such as it degrade user investigate knowledge, it assists transmission of virus in network, it increases load on network traffic. It also misuses user time, and energy for legal emails among the spam. For evade spam there are so many conventional anti-spam technique includes Bayesian based sort, rule based system, IP blacklist, Heuristic based filter, White list and DNS black holes. These methods are based on satisfied of the post or links of the mail. In this paper we proposed an efficient spam filtering technique based on neural network. The technique used is RBF a neural network technique in which neuron are trained. The results obtained by using this technique are compared with SVM. The parameter meter for comparison is precision and accuracy. On the basis of these two parameters we compared the proposed technique with SVM.

Index Terms—AdaBoost, Content Spam, Black and White Listing, Link Learning, RBF, Spam Filter, SVM.

I. Introduction

Spam is termed as unwanted money oriented mail [4]. It is needed to note down that certain individuals want to have this type of messages. There are viewers for e-mail publicity, regardless of the product that is being sold. Spammers try to reach these types of individuals. They do not know which people are made up of this group. They send spam to peoples, so that than can reach the people who comprise the group. This is because they don't know who will respond to the message and who will not. Spammers are the persons which are technically skilled and are hired by companies to send spam. Through a third party, the companies try to keep themselves from take legal action [5]. Spamming can be very profitable for a company if it done in right way. Let's take an example company is selling defected toys for 50 dollars a toy. If the company lets the spammer send out 10 million mails and the response rate is just 0.1% it will make half a million dollars [3].

E-mail addresses are get by the spammers through websites, newsgroups etc. [6]. It can be turn this into a benefit, by fooling spammers with foggy e-mail addresses and thus collecting their spam.

Spam is not restricted to e-mail. It exists in text messaging services (SMS) [8], newspapers and other communication media. SMS spam can cost even more than E-mail spam. For example, user has subscribed to receive a notification via SMS when they receive e-mail at their mail account. They pay for every SMS received regardless if it is a spam or a ham.

Cell phone spam is a type of junk message in the form of text message. This is defined as SMS spam or text message spam.

As the vogue of Cell phones rushed in the late 1990s, consistent users of mobile began to see a large scale growth in the number of unwanted marketable advertisements being sent to their cell phones through text messaging. This is mostly irritating for the receiver because, unlike in email, some receivers may be charged a fee for every spam message.

Cell phone spam is usually less tenacious than mail spam, where in 2011 around 95% of email is spam. The volume of mobile spam differs generally from area to area. In America, SMS spam has increase at large scale from 2007 through 2013, but remains below 1% as of December 2012. In Asia approximately 35% of sms were spams in 2013[32].

II. Literature Survry

Ramachandran A et al .in their work they studied the network level junk mails. The spams are detected from the network level. In their work they use DNS server for hosting. Spams are detected at IP level. BGP router algorithm is used to detect the spam mails [7].

Krishnan et al. proposed an Anti- trust Rank algorithm for the web spam detection. The algorithm is based on the approximate isolation principle. Threshold values are set for the set of spam web pages [13]. The results obtained from Anti Trust Rank and Trust Rank algorithms are compared.

Carlos Castillo et al presented a learning algorithm WITCH (Web Spam Identification through Content and Hyperlinks). This algorithm during learning phase uses the hyperlink structure in addition to page features. The way of graph regularization is used to utilize the hyperlink which yields a predictor that differs smoothly among interconnected pages [13].

Guang Gang Geng et al. in their work they proposed a semi supervised learning link based algorithms. These algorithms are used to speed up the performance of a classifier. This classifier merges the old self-training with topological dependency based on link learning [12]. The Experiments with WEBSPAM –Uk20006 (http://chato.cl/webspam/datasets/uk2006/) benchmark indicated that the algorithms are productive.

Loredana Firte et al. presented a new approach for spam detection filter. The solution proposed is an offline application that uses the K-Nearest Neighbor algorithm and pre- classified email data set for the learning process. KNN algorithm classified the messages which is based on features extraction from the email‘s properties and content[18].

Rafiqul Islam and Yang Xiang did sorting of worker emails form saturation of spam. In their paper, “Email Classification is done with the help of Data Reduction Method” which is an effective email classification method. This method is based on data purifying technique. A novel filtering technique using instance selection method (ISM) is introduced. ISM reduces the useless data samples from training model and then categorizes the test data. ISM helps to recognize which samples (examples, patterns) in email should be selected as representatives of the all dataset, without any loss of facts. They have used WEKA tool in our joined classification model and tested diverse classification algorithms [23]. Their experimental studies illustrate significant performance in terms of classification accuracy with reduction of false positive instances.

M.Basavaraju et.al, in this paper text clustering spam detection approach which based on vector space model is proposed. By using this technique one can detect email as spam and non- spam. The Proposed method contains the distance among all of the elements of an email [19].

Saadat Nazirova performed a work,” Survey on Spam Filtering Techniques”. In this paper the existing e-mail spam filtering methods are described. The grouping, judgment, and comparison of traditional and learning-based techniques are provided. Some private anti-spam products are verified and compared [26]. The declaration for new method in spam filtering technique is considered.

Faraz Ahmed et al. [9] Markov clustering based approach for the detection of spam profiles on OSN’s is presented in this paper. Work is based on a data set of Facebook profiles, which include both benign and fake profiles. Three features are identified and used for to model public collaboration of OSN user using a weighted graph. Markov clustering is applied to exploit the behavior similarity of profiles and mine the cluster existing in profile data set.

R. Kishore Kumar et al proposed the survey of email spam filter over data mining techniques. In their work, “Comparative Study on Email Spam Classifier using Data Mining Techniques” is proposed. TANAGRA data mining tool is used to analyze the spam data .It explore the efficient classifier for email spam classification. Firstly, feature creation and feature selection is done to draw out the relevant features. Then numerous grouping algorithms are applied on this dataset and cross validation is done for each of these classifiers [25]. In conclusion, best classifier for email spam is acknowledged on the basis of error rate, precision and recall.

Lourdes Araujo et al, present a work in which they tries to detect the tweets as spam in actual time by means of language as primary tool. Paper also introduced an general valuation method that has permitted showing how the system is able to obtain an F-measure at the same level as the best state of the art system based on the detection of spam accounts [16].

Siddu.Pacingill. Algur et.al, proposed a system in which spam web pages are detected with the help of link and content spam detector. System classifies the web page as spam based on threshold which is set by algebraic method. Unsupervised web spam detection problem is studied. For link spam detection the URL is taken as target and the link spamicity is calculated. In content spamicity the content of the web page is considered [29]. The result obtained from both is average spam score which is compare with the threshold value?

Nosseir, Khaled Nagati and Islam Taj-Eddin performed a work,” Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks”. They proposed a character-based technique. A multi-neural networks classifier is used by this approach. A normalized weight values derived from the ASCII value of the word characters are used to train the neural network [3]. Results obtained from experiment show high false positive and low true negative percentages.

Sahil Puri et al, in this paper spam detection is done on the basis of content and a rule based filtering. A new filter has been introduced in suggested work by the interfacing of rule based filtering followed by content based filtering for more efficient results [27].

Mohammed Mikki et al [1] An improved filtering technique is presented which is based on the improved digest algorithm and DBSCAN clustering algorithm.

Vandana Jaswal in this paper an image spam detection system is introduced. Hidden markov model was used to detect all the spam images.

Asmeeta Mali [6] performed a work, “Spam Detection using Bayesian with Pattern Discovery”. In her paper she proposed an operative procedure to recover the efficiency of using and apprising revealed patterns for conclusion appropriate information using Bayesian filtering algorithm and effective pattern. Discovery technique we can detect the spam mails from the email dataset with good correctness of term.

Neha Singh performed a work, “Dendritic Cell algorithm and Dempster Belief Theory Using Improved Intrusion Detection System”. To reduce the false alarm rate she proposed a new dual detection of IDS based on Simulated System that assimilating the Dendrite Cell Algorithm and Dempster Belief theory in her work [21].

R.Malarvizhi et al.a summary for spam filtering, and the techniques of evaluation and evaluation of different filtering methods is present in this paper. Fisher Robinson Inverse chi square, Ad boosts classifier, Bayesian classifiers are discussed. Bayessian method is used to create the spam filter in this paper [24].

III. Proposed Solution

From literature survey we have studied the various techniques which are used for the spam detection. These techniques have the problem with accuracy and precision. We proposed the RBF technique. It is a neural network technique. Proposed technique improves the accuracy and precision. Results obtained from the RBF are compared with the SVM.

IV. Work Done

Step1: Spam can be identified by various methods. From literature survey we studied the various techniques which are used to detect the email as spam.

Different spam detection techniques are considered. These techniques are found to have some limitations.

Step2: Markov clustering, DBSCAN, List filtering are available techniques which are used to detect the spam. These techniques have less accuracy, precision values.

After the spam detection accuracy precision and

Step3. Pervious methods have less accuracy and precision. We proposed the RBF which produces the better output. It is a neural network based technique. Another technique named SVM is used. We compare the result which we obtained from both the techniques.

Step4: All experiments are performed in MATLAB framework. The framework is used for the implementation of the SVM and RBF algorithms. The RBF algorithm is implemented to detect the email as a spam or ham. We calculated the accuracy, precision. The

Step5: Implementation: The implementation of proposed technique is described. Firstly we identify the problem, then we start to find the solution of this problem which we describe step by step in the diagram. Matlab is a framework where we done the implementation.

For spam detection firstly we collect the spam words. We create a spam word dictionary. These words are used for training and testing. After the creation of dictionary we need to extract the feature of these words so that we can use these words for training and testing.

Feature extraction can be done by various ways. In previous feature extraction is done with the help of clustering. We are doing the feature extraction on the basis of weights of the alphabets. The process of feature extraction is done as follow.

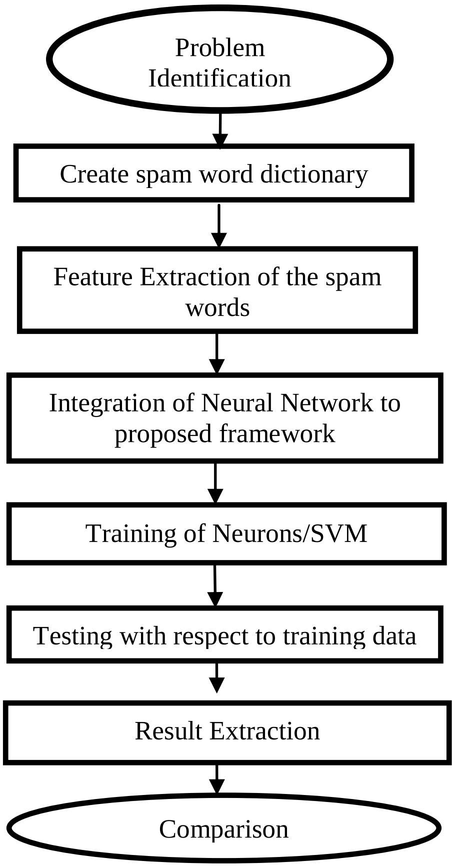
Problem Identification
Create spam word dictionary
Feature Extraction of the spam words
Integration of Neural Network to proposed framework
Training of Neurons/SVM
Testing with respect to training data
Result Extraction
Comparison

Table 1. Pseudo Code for feature Extraction

```
Step1   Load_ all = input all words
Step2   for each word in load _ all
          Value(i) = generator(word);
Step3   End
Step4   Void generator()
Step5   for each char in word  S= searchpos(g. file)
          If S! = Empty
          Var val = Val + S;
```

After the feature extraction the training is done using SVM and RBF. RBF is a neural network technique which uses the hidden neurons to process the input and to give the output.

SVM create a hyper plane which separate the different types of data. The training of SVM is done by following these steps

Table 2. Pseudo code for SVM

```
Step1 Initialize training data ($x_i$, $y_i$) for i = 1…N
Step2  Generate the weight vector and bias such that
       $f(x) = W^t + b$
Step3  Train data using  svmtrain
Step4  Generate the groups for training set such that
       svmstuct = svmtrain(training set, groups)
Step5  Find the support vector by using svmclassify)
       such that svmclassify(svmstruct, testdata,groups
Step6  end
```

RBF training and testing is done. The Liebenberg algorithm is used in this technique. Following is the pseudo code for RBF

Table 3. Pseudo Code for RBF

```
Step1  For each word set in all word generate weight
     = $ax + b$
       Where a = constant, x= provided data, b= random
       weight
Step2   Epoch. System = 100;
Step3   Hidden Neurons = 10
Step4  Initialization fn = init p;
Step5  Training function = trainln
Step6   Type = “ feed forward back propagation”
        Method = “RBF”;
        Algo = “Lavenberg”;
Step7   If processed hidden neuron == true;
Step8   Find epochs
Step9   Error;
Step10  Output layer = sim(train set, test set)
              end
```

To compare the both techniques we calculate the accuracy, precision, recall, frr and far so that we can identify which technique is better. The values calculated are stored in a table. The according to these values are plotted.

Error is calculated firstly so that we can calculate the other values by using this. It can be calculated as:

Error= (training data – testing Data)^2/ Length of testing set

Now with the help of Error we can calculate our other values which we are used for result.

Far (False Acceptance Ratio): Number of spam classified as non- spam. It also called false positive ratio.

Far= (error- test data) / Length of testing set.

Frr (False Rejection ratio): Number of non-spam classified as spam. It can be calculated as:

Frr= (Error – Far)/ Length of testing set.

Precision: it is the percent of positive spam data that is correct. We can calculate it as:

Precision = (Test set-Error)/ Length of testing set.

Recall: Its value should be low. It is percentage of positive labeled instance.

Recall= (Test set-precision)/ Length of testing set.

Accuracy: It describes how close a measured value to actual value. The technique which has high accuracy is better.

Accuracy= (1-(far + frr)/100)

By using above formulas we calculate the accuracy, recall, precision, Frr and far. Now with the help of these values we identify which technique is better for the spam detection. Values are listed in tables. Below table 1.4 contains the values which we obtained by using RBF and table 5 contains the values which we obtained by using SVM.

Table 4. RBF values

File	Iterations	Accuracy	Precision	Recall	Frr	Far
Testset1	4	99.8087	0.0043073	4.7144	0.000787	0.19048
Test Cat1	5	99.7879	0.0015895	14.392	0.003149	0.20894
Test Cat2	8	99.9546	0.0043407	43.1928	0.002124	0.04329
TestsetCat3	5	99.7846	0.003895	0.82595	4.147e-05	0.2157
Test Cat4	6	99.8879	0.004307	0.3569	7.9753e-005	0.11203

Using RBF we calculated the above values. These values are used to plot the graphs.

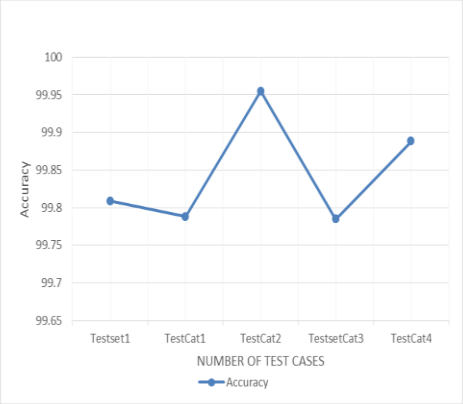

Fig.1. Accuracy using RBF

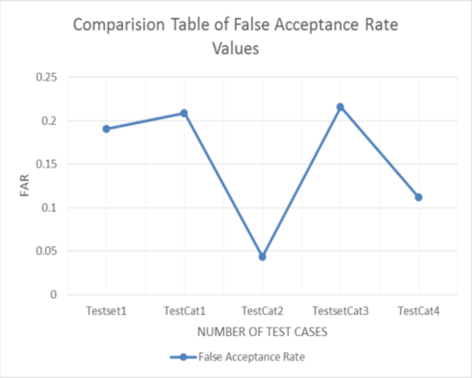

Fig.2. Recall using RBF

Recall is plotted against diiferent testsets. Recall value shuold be low.

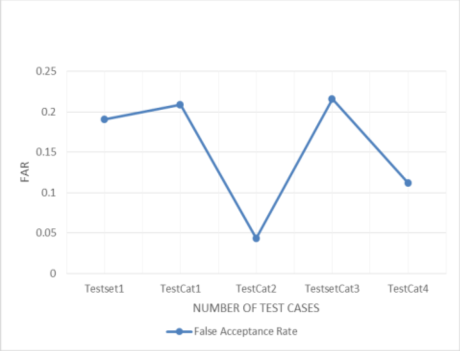

Fig.3. Far using RBF

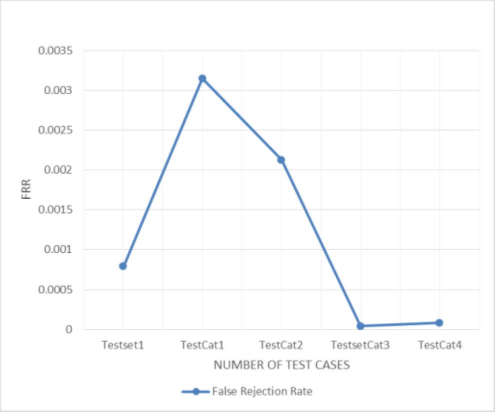

Fig.4. Frr using RBF

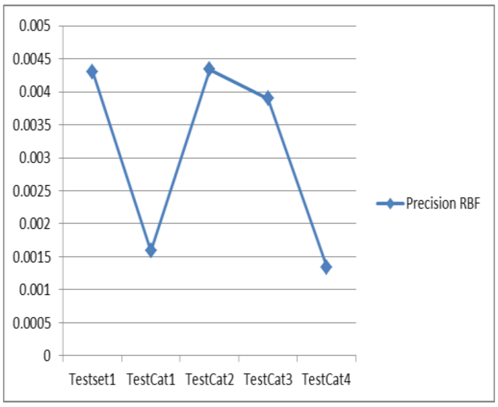

Fig.5. Precision using RBF

Table 5. SVM values

File	Accuracy	Precision	Recall	Frr	Far
Textset 1	88.9377	0.0034517	0.30082	0.0035667	0.10706
TextCat 1	88.9135	0.001225	0.28009	0.00012264	0.11074
TestCat 2	88.9377	0.0034517	0.30082	0.003567	0.10706
TestCat 3	88.9376	0.0034517	0.30082	0.0035667	0.10706
TestCat 4	88.9345	0.001225	0.30056	0.000123	0.1174

Using SVM we calculated the above values. Graphs are plotted using these values.

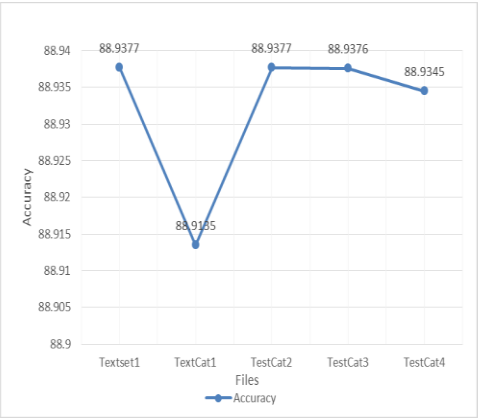

Fig.6. Accuracy using SVM

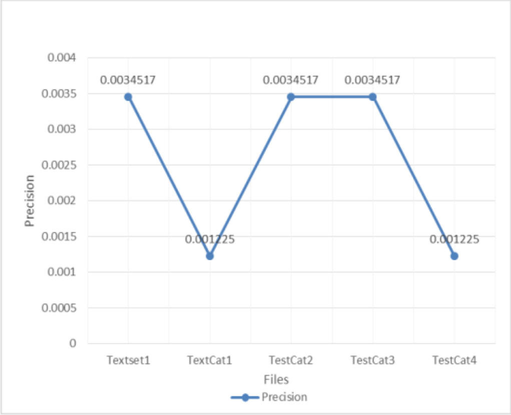

Fig.7. Precision using SVM

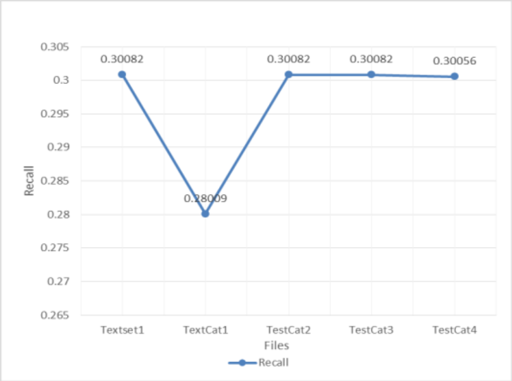

Fig.8. Recall using SVM

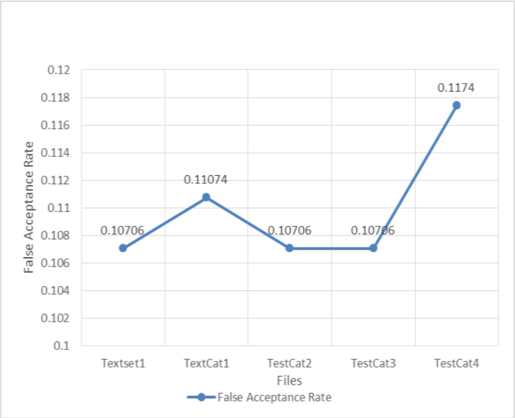

Fig.9. Far using SVM

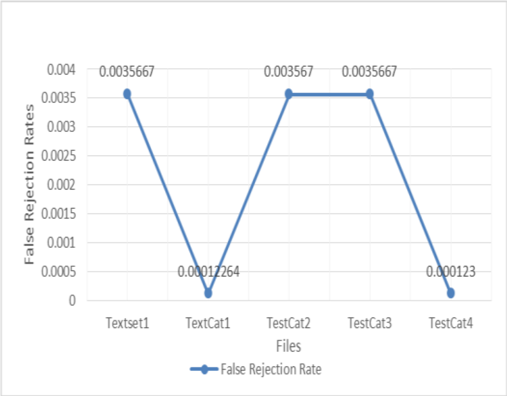

Fig.10. Frr SVM

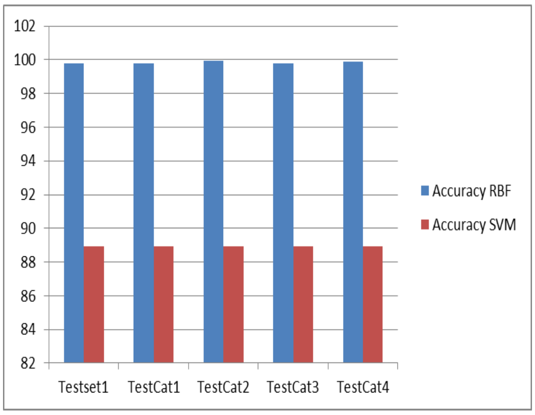

Fig.11. Comparison of Accuracy

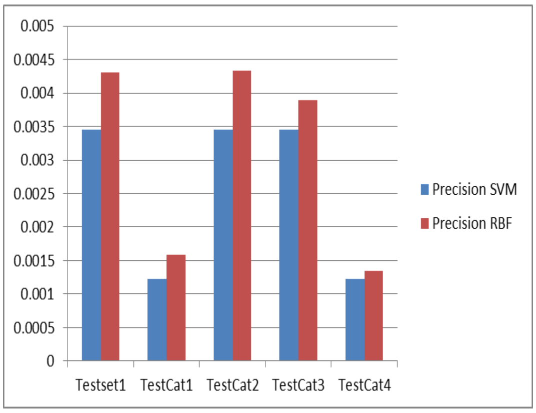

Fig.12. Comparison of Precision

The above graphs show the comparison of accuracy and precision. The accuracy and precision values of RBF are higher than the SVM. The proposed technique gives the better results.

V. Conclusin And Future Scope

During the study of this dissertation, it has been widely observed that there are numerous spam detection techniques available around us. Most of these techniques either lack in performance or level of accuracy. The proposed methodology is adopted to enhance the precision quotient of the existing spam detection methods. New mechanism using RBF is proposed. The proposed mechanism improves the accuracy, precision, recall Frr and Far. The proposed mechanism is compared with SVM and the results have been comparatively better. We use a database of approximately 1000 spam words in our current research work; in future we can use larger data set for spam detection. The advanced neural network techniques can be used in future for better results. The proposed algorithms can be used with other algorithms to make a hybrid algorithm which helps to improve the performance of the spam detection system.

References

[1] Gomes, L.H, Caztia , in *Proceeding 4th ACM SIGCOMM Conference on Internet Measurement ,ACM*, pp.356-369, 2014

[2] Megha Rathi and Vikas Pareek, "Spam Mail Detection Through Data Mining –A Comparative Performance Analysis", *International Journal of Modern Education and Computer Science*, Vol.5, No.5, 12 December 2013.

[3] R.Malarvizhi and K. Saraswathi, "Content – Based Spam Filtering and Detection Algorithms-An Efficient Analysis and Comparison", *International Journal of Engineering Trends and Technology*, Vol.4, Issue 9,Septmber 2013

[4] N. Singh," Dendritic Cell Algorithm and Dempster Belief Theory Using Improved Intrusion Detection System "*International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, Issue 7, July 2013 ISSN: 2277 128X

[5] Asmeeta Mali, "Spam Detection Using Bayesian with Pattern Discovery", *International Journal of Recent Technology and Engineering*, ISSN: 2277-3878, Vol.2, Issue-3, July 2013.

[6] J. Vandana and N. Sood, "Spam Detection System Using Hidden Markov Model", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.3, Issue 7, July-2013

[7] Ahmed, H. Alaa, "Improved Spam Detection using DBSCAN and Advanced Digest Algorithm", *International Journal of Computer Applications*, Vol. 6, May 2013.

[8] S. Puri, M. Ahuja, "Comparison and Analysis of Spam Detection Algorithms", *International Journal of Application or Innovation in Engineering and Management*, Vol. 2, Issue 4, April 2013

[9] A. Nossier, khaled kagati, and Islam Taj-Eddin, "Intelligent Word- Based Spam Filter Detection Using Multi Neural Networks", *International Journal of Computer Science*, Vol. 10, Issues 2, No. 1, March 2013 ISSN(Print): 1694 -0814|ISSN(Online): 1664- 0784

[10] S. Roy, A. Patra, and S. Sau, "An Efficient Spam Filtering Techniques for Email Account", *American Journal of Research*, Vol. 2, Issue 10, 2013

[11] F. Ahmed and M .Abulaish, "An MCL Based Approach for Spam Profile Detection in Social Network", in *proceedings of 11th International Conference on Trust, Security and Privacy, IEEE*, pp. 602-608, 2012

[12] T. Mahmoud and M.Mahfauz, "SMS Spam filtering Techniques Based on Artificial Immune System", *International Journal of Computer Science*, Vol. 9,Issue 2, No.1, March 2012

[13] J. Martinez Romo and Lourdes Araujo, "Detecting malicious Tweets in Trending Topics Using a Statistical Analysis of Language", *Elsevier*, 2012

[14] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, "Comparative Study on Email Spam Classifier using Data Mining Techniques", in *Proceedings of the International Multi Conference of Engineering and Computer Science*, Vol 1, 2012

[15] Siddu. P, Alger and N. tarannumPendari, "Hybrid Spamicity Approach to web Spam Detection", in *Proceeding of Conference on Pattern Recognition, Informatics and Medical Engineering IEE*, March 2012

[16] Tiago. A "Contribution to the study of SMS Spam Filtering New Collections and Result", in *Proceedings of the 11th ACM symposium on Document engineering ACM*, pp. 259-262, Sept 2011

[17] S. Nazirova, "Survey on Spam Filtering Techniques", *Communication and Network*, August 2011

[18] G. Kumari Tak and S. Taposwi, "Query Based Approach towards Spam Attacks Using Artificial Neural Network", *International Journal of Artificial Intelligence and Application*, No.4,Oct 2010

[19] M.Basavaraju and Dr.R.Prahakar, “A Novel Method of Spam Mail Detection using Text Based Clustering Approach”, *International journal of Computer Applications*, Vol. 5, August 2010.

[20] R. Islam and Yang Xiang, “Email Classification Using Data Reduction Method”, in *Proceeding of 5th International ICST Conference on Communications and Networking in China (CHINACOM), , pp. 1-5. IEEE, 2010, June 16, 2010*

[21] L. Firte, C. Lemnaru and R. Potolea, “Spam Detection Filter Using KNN Algorithm and Resampling”, in *Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing*, pp. 27-33. IEEE, 2010.

[22] M.Soranamageswari and C.Meena, “Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Network”, in *Proceeding of 2nd International Conference on Machine Learning and Computing*, February 2010

[23] Guang-Gang Geng, Qiu-Dan Li and Xin-Chang Zhang, “Link Based Small Sample Learning for Web Spam Detection”, *ACM*, April 2009

[24] J. Abernethy, Oliver Chappelle and Carlos Castillo, “WITCH: A New Approach to Web Spam Detection”, in *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIR Web*. 2008)

[25] Alex Brodsky (Canada) and Dmitry Brodsky (USA), “A Distributed Content Independent Method for Spam Detection” *International Journal of Computer Applications*, 2007.

[26] A. Ramachandran and Nick Feamster, “Understanding Network – Level Behaviour of Spammers”, *ACM*, September 2006.

[27] Sender policy framework *(SPF)* for authorizing use of domains in e-mail, Version.1, No. RFC 4408. 2006.

[28] Erik D. Demaine, F.H. Meyer auf der Heide, U. Paderborn, Rasmus Pagh, Mihai Pˇatra¸scu, “On Dynamic Dictionaries Using Little Space”, *ARVIX*, 2005.

[29] P.A.Chitira, J.Diederich and W.Nejdl, “MailRank: using Ranking for Spam Detection”, in *Proceedings of 14th International Conference on Information and Knowledge Management ACM*, October 2005.

[30] Anti-Spam site, Claws and paws, Aug 2004 [Online]. Available: http://www.claws-andhttp://www.claws-andpaws. com/spam-l/tracking.html

[31] J. Junod, “Serves to Spam: Drop Dead”, *Computer and Security Elsevier*, Vol.16, 1997

[32] Anti - Spam abuse site – http://spam.abuse.net/

Authors’ Profiles

Reena Sharma student at Chandigarh University. My area of interest is Neural Networks**.** I have completed my thesis work on Email spam detection.

Er. Gurjot kaur Assistant Professor at UIE-CSE Chandigarh University. I have completed my M.tech from Punjabi University Patiala. My area of interest cloud computing, network security and Cryptography.