Spam Detection Using ICNEURO for Enhanced Accuracy

Robin Sharma
M.Tech, Computer Sci. Engg,
RIMT-IET, Mandi Gobindgarh,
Punjab, India

Principal Dr. Sushil Garg
RIMT-MAEC, Mandigobindgarh
Punjab, India

***Abstract*— Spam detection or filtering process is required to cope with the harmfull effect of spam e-mails affecting directly or indirectly to the users. SPAM e-mails have a direct cost in terms of time, server storage space, network bandwidth consumptions and indirect costs to protect privacy and security breaches [6]. For providing solution to solve this problem various techniques has been implemented. This paper present the ICNEURO which is formed with the combination of Independent component analysis and Neural Network technique for enhancing the accuracy of spam detection from the dataset which is basically in the textual form applying the content based filtering technique. We make an ICNEURO as a user level program which uses the advance feature of Independent Component Analysis (ICA). Results of our approach show the enhancement in accuracy as the content or words will increase.**

Keywords- ***Spam; Neural Network; content based filtering; Independent Component Analysis; Principal Component Analysis component***

I. INTRODUCTION

SPAM is defined as an unwanted of electronic message [6] posted blindly to thousands of recipients [12] [6] also known as ‘junk e-mails’ or ‘unsolicited bulk email’ (UBE). Receiving of unwanted emails in account which are not even requested nor subscribe would consider as spam emails. The electronic messages has mainly used to provide information or communicate with receiver only in case if ever user demand for information, had communicate in past or receiving information relevant to its interest. But getting something out of the box could have good or bad effect depend on the sender aim. The mail may be informative which send with a purpose of providing valuable information or with harm full effects purpose of breaching security or privacy or anyhow create trouble.

Spam is undesirable because it eats up resources like disk space and user time. It has a regress effect in terms of time, money, storage space and faulty effect on security, threat of stolen data from system and may more. For example, financial theft, identity theft, data and intellectual property theft, virus and other malware infection, child pornography, fraud, and deceptive marketing are usually caused by SPAM messages that point to links to collect personal information, open porn websites, or download virus [6]. Many commercial organisation, advertiser, promoters use the internet for spreading information through emails to their clients or to other user but the spammer always came with a new idea to breach every prevailing privacy level. Spammers are not only threat to recipients but they are also affecting the business of organisations. In many countries after realization of the impact of SPAM, a different laws and legislations have enacted but this is not the only way to fight against spammers.

There are number of different approaches used to filter the received messages and made a measure to protect our accounts from SPAM. Several technical solutions like commercial and open-source products have been used to alleviate the effect of this issue. Every approach performance is measured in terms of its false positives and true positives.

II. BRIEF STRUCTURE OF EMAIL

An electronic mail structure made with three vital components: the envelope, the body of message and the header and other is IP

A. Body

The main content of message which we always see and additionally some of the content are highlighted, bold etc.

B. Header

This part contains routing information, including sender and recipient, date and subject. Email address has two parts separated by ‘@’.Username on the left side and domain name for the host server at right side.

C. Envelope

This part is hidden from user as it contains the internal process routing in formation.

D. IP Address

It contains the sender Internet Protocol address.

III. LITERATURE REIVEW

Keeping in view the adverse effect of SPAM and essential methods require filtering messages, so that they can be classified between SPAM and Non SPAM messages. This section presents the review of papers, which were based on some of technical anti-spam approaches.

A. *User defined filters*

In this approach user can control the filtration process. They can form their own rules on basis of which messages passed and decide which one is spam and others not.

B. *Header filters*

Every email has a header part provide the sender email address. This filter analyses to detect fake header through matching list of address which users have in its list of receiving address.

C. *Content filters*

These scan the body of the email message and use the presence, absence and frequency of words to determine whether the message is spam or not. Content based filtering has been found to be most effective for spam detection [3].

Abhimanyu Lad [3] describe a spam detection user level program called spamnet which use the heuristic rules, artificial neural network as classifier and most important principal component analysis to detect spam by analysis mail content. The detection process is automated which retrains its system every 7 days so that it adapt the changes as new mails pattern changes. The process first pre-process the message through extractor module whose work is vital in whole process. Its main functionality is to remove of stop words and parses the message according to [3] rules. This phase is also responsible in providing input to PCA which are the outputs of extractor. The input to PCA is feature vector which are created from words with the help of volatile vocabulary. The main role of PCA is to transform feature vector to optimal representation which has done by computing eigenvector and improve the performance and efficiency by reducing the feature set. And in final neural network with 6 inputs classify the mails through its output signal spam or non-spam.

Alex Brodsky et al. [4] proposed distributed, content independent, spam classification system called TRINITY that is specifically aimed at botnet generated spam and can be used in combination with existing spam classifiers [4]. They mainly focus on method which protect from the botnets attack. The computers which are connected to internet are first hacked and from their system IP address which are assigned dynamically at boot time, a numerous junk emails send to users. Trinity is a distributed spam detection system [4] whose main objective along with detection process is to keep some of points in consideration while developing this approach are that it should be easily installable ,pluggable, within existing infrastructure. No need to modify the existing protocols. With the use of distributed system, central point of failure clause removed and most vital component it would be user controllable. The main goal of this approach is to identify the source an email which sends the spam and immediately update the distributed database about the sender information so that whole system disable the breach.

Dominic Langlois et al. [5] presents two Independent component analysis (ICA) algorithm which are Infomax and FastICA algorithms. With the implementation of ICA algorithm the results were also compared between them and also along with principal component analysis. By taking 'cock tail party effect' the result obtained by PCA and ICA are compared which state that for gaussianity value has to be assumed for PCA and for ICA non-gaussianity value whereas the PCA would not give better result as ICA give.

ICA is use to represents a linear combination of the original variables or non-gaussian data so that the components would be statistically independent. The main goal of this paper is to find the independent component. ICA is a technique which used for source signal extraction but there implies some condition which has to follow for effective use [5]. To accomplish this task mixture of three images has taken as example and the algorithm which extracts components from the mixture image would prove is more efficient than other.

A. Nosseir et al. [6] present approach which is based on character-word based technique and for classifying multiple neural networks used. The content based filtering technique check the words which consist 3, 4 and 5 character. After stop-word and noise word removal, the stemming process changes the plural form of nouns to root word. The list of words then classified according to the length of and categorizes them as good and bad words. Then for each word according to different length ,neural network has been trained which then classify which words in message are good and bad by giving output 0 and 1 respectively. Through this approach it has been proved that in comparison with the white list and black list this approach according to word length filtering perform better and along with it trains it sub network for improved performance.

IV. PROBLEM DEFINATION AND FORMULATION

Spams are the textual context of the system which can damage our system. Our basic problem is to protect our system from such unwanted files. To save our system form such kind of failures we need to design a system which can recognize the spams and can let you know on the basis of a training system. There are various spam detection techniques and algorithm available where each has its own advantage and disadvantage.

In previous latest work, Principal Component Analysis has been used with neural networks for spam detection. PCA is used before the implementation of classifier and it takes feature vector as input and after applying dimensionality reduction technique PCA reduce the number of inputs which then further forwarded to neural networks. As a classifier, neural network is able to detect spam efficiently. But PCA takes eigenvectors that are highly correlated to each other.

To remove inefficiencies created by PCA, FastICA can be used. Independent component analysis (ICA) is a method whose function is to find independent component data. ICA is a statistical method that expresses a set of multidimensional observations as a combination of unknown latent variables. It works in non-Gaussian values or data. As single technique may become able to detect spam but according to the evaluation result their performance may not be appreciable so to overcome with that thing, a combination of one or more technique can be applied for the detection process.

In this research, FastICA will be used for spam detection with neural networks. FastICA is signal processing technique used for analysis of several types of data and feature extraction. Using Blacklist and Whitelist approach, a content based technique filter the message. With the use of ICA algorithm these input files are then changed into Digital signal for pattern matching. Then Neural Network by taking converted signal as input will classify which mails are SPAMS and which are NON-SPAM.

V. Motivation and Objective

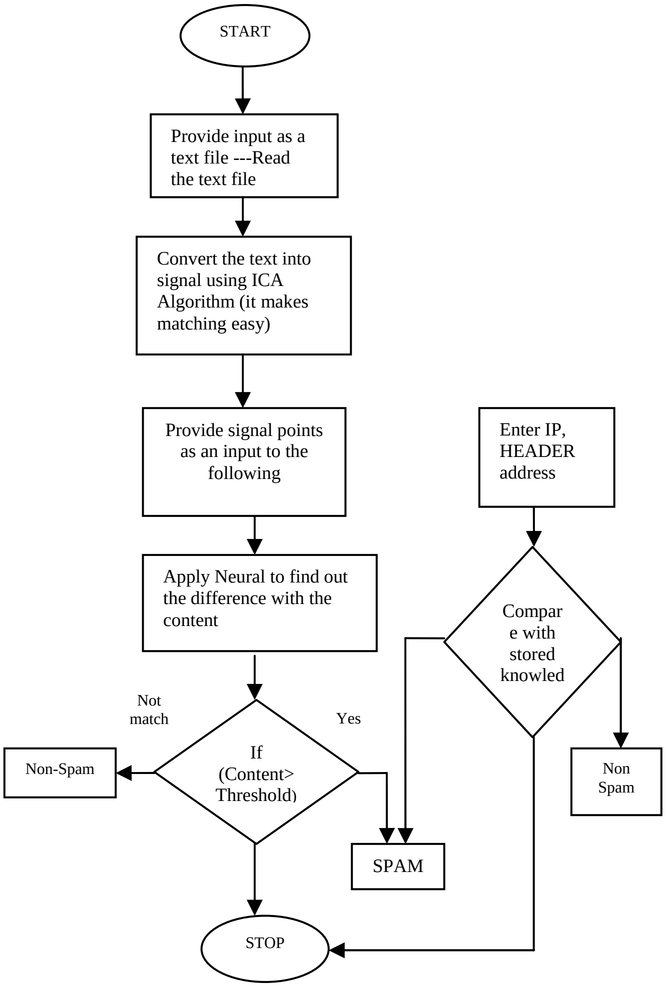
Keeping in view the adverse effect of SPAM and essential methods require in filtering messages, so that they can be classified between SPAM and Non SPAM messages. The aim of this thesis is to design and implement the approach which gives more accurate results than previous work done on spam filtering. This project applied Independent component analysis (ICA) with Neural Network to detect spam mails. The algorithm of Independent component analysis is used to convert the text into digital signal whereas neural network as classifier gives the output by taking input as signal and giving output by classifying e-mails as SPAM or NON-SPAM. ICNEURO, a combination of ICA and Neural network is a user level program, i.e. it runs as part of the user mail client rather than sitting on the mail server providing general services to its users. In order to achieve this aim, the following objectives must be fulfilled:

1) To design a system for spam detection
2) To train the system about spam
3) To check the documents on the basis of stemming and pattern matching
4) To increase the accuracy of the spam detection.

VI. Proposed Model

For classifying mail we developed a user friendly interface system called ICNEURO by using MATLAB tool. The input to the system would be textual files which contain the e-mail content and that textual file converted into digital signal form with the implementation of ICA. The output of ICA is next input to the Neural Network which further compared with selected file which contain the information about spam words. At end the neural Network will decide to classify as spam or non-spam.

A. FlowChart of Model

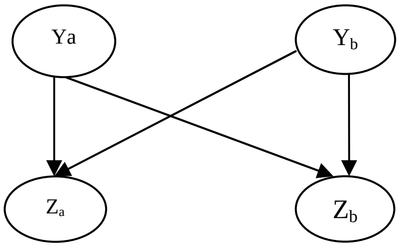

START
Provide input as a text file ---Read the text file
Convert the text into signal using ICA Algorithm (it makes matching easy)
Provide signal points as an input to the following
Enter IP, HEADER address
Apply Neural to find out the difference with the content
Compar e with stored knowled
Not match
Yes
If (Content> Threshold)
Non-Spam
SPAM
Non Spam
STOP

The logic behind comparison method is to divide every textual content file into two column parts and then comparison is performed with each subpart of other file for example if below presented line will be content of mail.

“This is the example which would help to understand the comparison logic”

Then division creates the:

	This	is	
	the	example	
Ya	which	would	Y_b
	help	to	
	understand	the	
	comparison	logic	



Ya and Y_b: Mail content

Z_a and Z_b: file from which comparison has to perform

Figure1. Division of Text

B. *Simulation of ICNEURO (using METLAB)*

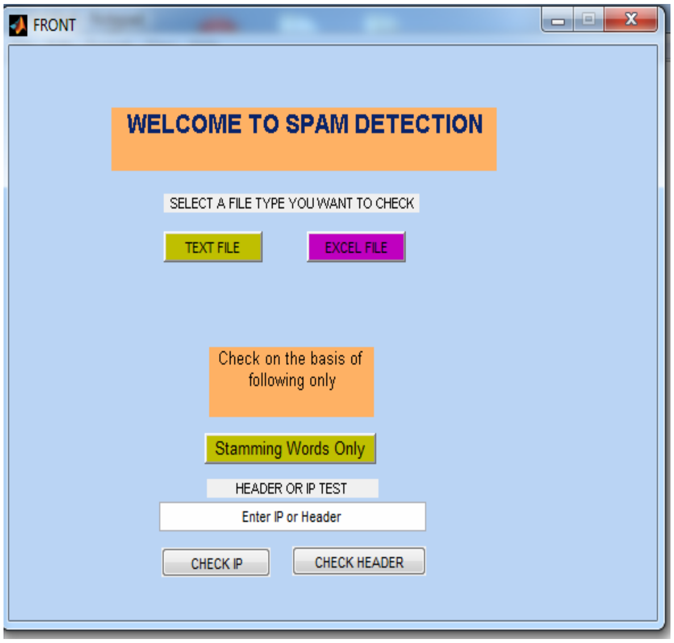


Figure2 Front page

Fig.2 shows the front page which has number of buttons and come first through this page only we can run the simulation of ICNEURO. The TEXT FILE and EXCEL FILE button enable to upload the text file which has all the e-mail content through browsing dialog window.

Fig.3 shows that after selecting text file, ICA algorithm convert the textual representation into digital signal and a message box appear which states that our chosen file has been successfully uploaded. When we click on SELECT SPAM DETECTION FILE button then a browsing dialog box will open through which we have to select those text file from which comparison or pattern matching has to be performed for detection of spam.

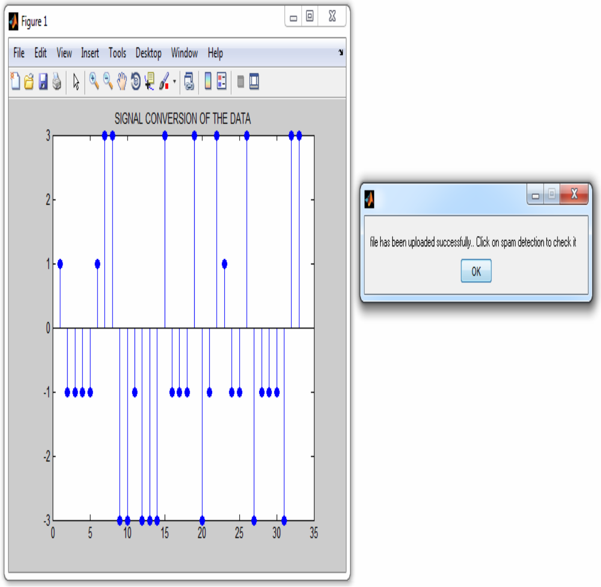


Figure 1
File Edit View Insert Tools Desktop Window Help
SIGNAL CONVERSION OF THE DATA
3
2
1
0
-1
-2
-3
0
5
10
15
20
25
30
35
OK

file has been uploaded successfully.. Click on spam detection to check it

Figure3 Signal conversion of data

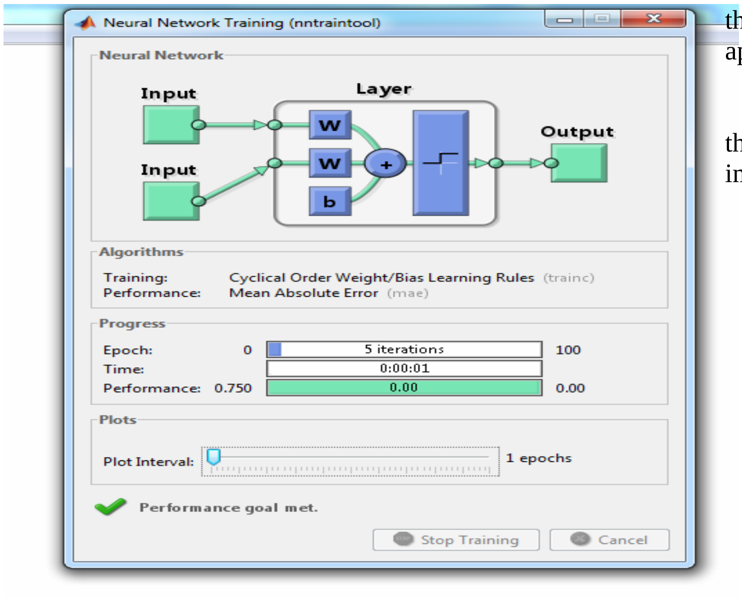
In Fig 4 according to our requirement we use the cyclical order incremental training function (trainc) because of its iteration procedure, this training function provide us output with better performance result.

1) *Number of Epoch:* An epoch is a measure of the number of times all of the training vectors are used once to update the weights. We set 5 iterations to Epochs value because generally 5 to 10 iterations are more than sufficient for neural network to decide what exactly the results analysis could be.

2) *Number of Inputs and Output:* We provide Two Inputs to the neural Network and One Output come from Neural Network which gives value between 1 and -1 where 1 indicates that mail is spam mail and -1 indicates that mail is not spam.

In Fig.4 the Neural Network generation diagram appear which gives the common detail about it and at end Fig.5 classifies whether mail is SPAM or NON-SPAM.

To check the header information we can also fill the header address in text edit button to confirm about spam header or not and similarly we can perform same for the Internet Protocol (IP) address.



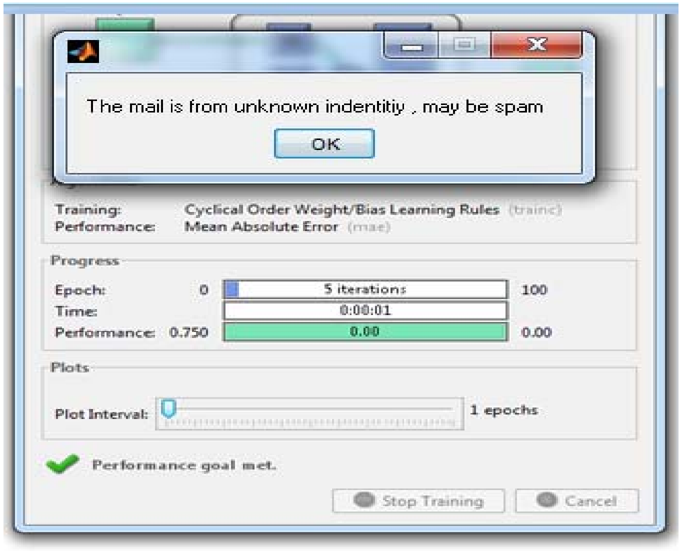
Algorithms
Training: Cyclical Order Weight/Bias Learning Rules (trainc)
Performance: Mean Absolute Error (mae)

Progress
Epoch: 0 5 iterations 100
Time: 0:00:01
Performance: 0.750 0.00 0.00

Plots
Plot Interval: 1 epochs

Performance goal met.

Figure4 Neural Network Training


The mail is from unknown indentitiy , may be spam

Training: Cyclical Order Weight/Bias Learning Rules (trainc)
Performance: Mean Absolute Error (mae)

Progress
Epoch: 0 5 iterations 100
Time: 0:00:01
Performance: 0.750 0.00 0.00

Plots
Plot Interval: 1 epochs

Performance goal met.

Figure5 Message shows the classification

The graph in Fig.6 shows the comparison between previous approach and our ICNEURO approach in term probability of error versus as number of word increase in file. Dataset that used in this project has taken form personal e-mail account and then the content of e-mail is saved in text file manually. In the graph X-axis shows the probability of error in percentage and Y-axis shows the File size words * 1000. For Example: if in Fig.6 along Y-axis it shows 10^{-1}, then it states that

1/10 *1000=100 words

This graph by comparing the result with previous approach states that if for 0 % error our approach requires 100 words then to show the same error rate the previous approach need approximately 110 words.

So this proves that as the number of words in a file increase the probability of error would decrease with the ICNEURO implementation.

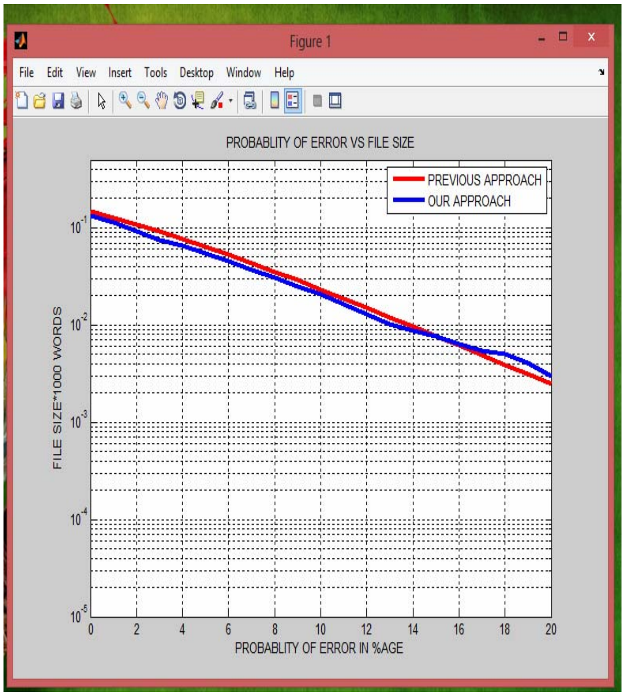


Figure6 Graph plot

VII. Conclusion and Future Work

With the work done we conclude that if ICA algorithm is combined with Neural Network it gives the one of the possible solution regarding the spam detection in terms of accuracy and computational complexity. The results are better as the ICA algorithm works on signal processing basically hence it is easy to find out the matching pattern or to compare the signal to another.

There are two sections:

1) Training Section
2) Testing section

Training section includes the Independent Component Analysis algorithm which converts the text into digital signal. The result of ICA algorithm is passed to the neural network. Then the neural network decides that how many iterations it would take to give you the best probabilities result generally 5 to 10 iterations are more than sufficient for neural network to decide what exactly the results analysis could be. We can also conclude that matrix to matrix mapping can provide batter solution rather than implementing into a single line system.

The only drawback within the current system has that it does not have any rule set generation pack for the processing. So if in future if somebody can combine fuzzy logic with neural network then drawback of the rule set will removed and the system would become much more efficient. For Experiment Neural Network classification can also be replaced with Bayesian classification to compare the performance of both the classifier.

ACKNOWLEDGMENT

As a student I would like to thank my college professor of computer science for valuable insights and comments that helped to improve this paper. And this work was supported by my parents for their grand support and making the work successful.

REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] Grigorios Tzortzis and Aristidis Likas, "Deep Belief Networks for spam filtering", 19th IEEE International Conference on Tools with Artificial Intelligence, GR 45110, Ioannina Greece (2007)

[3] Gaurav Kumar Tak and Shashikala Tapaswi, "Query Based approach towards spam attacks using artificial neural network", International Journal of Artificial Intelligence & Applications, October 2010

[4] Alex Brodsky (Canada) and Dmitry Brodsky (USA), "A distributed content independent method for spam detection".

[5] A.Hyvarienen and E.Oja, "Independent Component Analysis and Applications, Neural Networks" 13(4-5):411-430, 2000

[6] Ann Nosseir , Khaled Nagati and Islam Taj-Eddin," Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks",IJCSI, Vol. 10, Issue 2, No 1, March 2013.

[7] Thamarai Subramaniam, Hamid A. Jalab and Alaa Y. Taqa, "Overview of textual anti-spam filtering techniques", International Journal of the Physical Sciences Vol. 5(12), pp. 1869-1882, 4 October, 2010.

[8] Ahmed Khorsi, "An Overview of Content-Based Spam Filtering Techniques", Informatica 31 (2007) 269-277.

[9] Jaber Karimpour, Ali A. Noroozi, Adeleh Abadi, "The Impact of Feature Selection on Web Spam Detection", I.J. Intelligent Systems and Applications, 2012, 9, 61-67.

[10] A. Wiehes, "Comparing Anti Spam Methods", Master Thesis, Master of Science in Information Security, Department of Computer Science and Media Technology, Gjøvik University College, 2005.

[11] S. Heron, "Technologies for spam detection", Network Security, pp. 11-15, Jan 2009.

[12] Abhimanyu Lad, "SpamNET Spam Detection using PCA and Neural Network"

[13] http://www.cis.legacy.ics.tkk.fi/apo/papers/IJCNN99_tutorial web/node32.html

[14] Dominic Langlois, Sylvain chartier and Dominique Gosselin, "An introduction to Independent Component Analysis: Infomax and FastICA Algorithm" (2010)

[15] Sasmita Kumari Behra (2009) "FastICA for blind source separation and its implementation", Rourkela

[16] Martin, Spam Filtering using Neural Networks, http://www.web.umr.edu/~bmartin/378Project/report.html

[17] British Computer Society. Available: http://www.bcs.org/server.php?show=conWebDoc.14617

[18] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[19] V.Zorkadis, M.Panayotou, D.A.Karas, "Improved Spam e-mail Filtering Based on Committee Machines and Information Theoretic Feature Extraction", Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July-Aug,2005

AUTHORS PROFILE

Robin Sharma is persuing her Master's in Engineering in Computer Science from RIMT-IET, Mandi Gobindgarh, Punjab Technical University, Kapurthala. She is currently working on the project of Spam Detection for her research work. She has interests in subject areas like Pattern Matching, data mining, software engineering, web development etc.