ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

RSOAP: RBF Network Based Personalised Spam filter

Vandana S. Ahire [1]

Research Scholar, Department of Computer Engineering, S. S. V. P. S's B. S. Deore College of Engineering, Dhule, Maharashtra, India[1]

ABSTRACT*:* Spam is unsolicited messages sent over the Internet which nobody wants or ever asks fort .There are many different techniques to develop a correct and user friendly spam filter. The Existing Bayesian classification is used as a probabilistic learning method, which analyses contents of mail to calculate its probability of being a Spam using individual characteristic of words in the mail. But Existing Bayesian spam filters are easily poisoned by clever spammers who avoid spam keywords and add many innocuous words in their emails. Hence new improved method has been introduced called (RSOAP) RBF network based personalised spam filter. RBF network is an artificial neural network that uses radial basis functions as activation functions. It focuses on the impact of the widths of the Gaussian functions on accuracy of estimates generated by RBFN network The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. In this existing Bayesian spam filtering methods will be compared with RBF Neural networks is given.

KEYWORDS*:* Spam filtering, social networks, Bayesian spam filters, RBF Network.

I. INTRODUCTION

Spamming is the abuse of electronic messaging systems to send unsolicited bulk messages or to promote products or services, which are almost universally undesired. The Problem of Spam is currently of serious and escalating concern, and it is challenging to develop spam filters that can effectively eliminate the increasing volumes of unwanted mails automatically before they enter a user's mailbox. Spam emails interfere with both email service providers and end users. A fundamental way to prevent spam is to make it unprofitable to send spam emails, thereby destroying the spammers' underlying business model. The email spam consume a large number of server resources and bring great harms to the user's normal use.spam is unsolicited email message which is forced on people who would not otherwise choose to receive it. Spam is a combination of: unsolicited commercial e-mail (UCE) and unsolicited bulk e-mail (UBE) [6]. The three main characteristics of spam email are; (i) not requested by the recipients; (ii) has commercial value; (iii) always sent in bulk. Spam is a problem because it wastes the email storage, email recipients time to open, read and delete the email. It targets individual users with direct mail messages [6].

To support against unsolicited e-mails there are many recent research presented with goal of efficient, accurate spam filtering. Few previous spam filters can meet the requirements of being user-friendly, attack-resilient, and personalized.

II. RELATED WORK

Previous spam filtering approaches have mainly divided into two categories: content-based and identity-based [1].

A. Content-Based Approaches

Content-based filtering approaches are based on the assumption and it reads the text in order to discover distinctive features which are used for sake of classifying a message. It is used to analyses the headers, subject, and body of an email message using feature (token) matching or statistic methods to determine whether it is spam or legitimate email.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Many content-based filtering methods have been proposed to filter spam from email. An important part in content based filtering is feature selection [1].

The primary approach of content-based spam filtering is the static keyword list [1], which however makes it easy for a spammer to evade filtering by tweaking the message. The second category approaches includes machine learning-based approaches such as Bayesian filters [2], decision trees [9], Support Vector Machines [9], Bayes Classifiers [2], [3], [4] and combinations of these techniques [1].

- ***Overview Of The Basic Bayesian Spam Filter:***

A Bayesian filter [2] has a list of keywords along with their probabilities to identify an email as a spam email or a legitimate email (HAM). The list is built by training the filter. During training, a user is given a pool of emails, and s/he will manually indicate whether each email is spam or not. We use $P(S)$ and $P(L)$to denote the probability that an email is a spam email and a legitimate email, respectively. The filter parses each email for spam keywords. It calculates the probabilities that a word *w* appears in a spam email and in a legitimate email, denoted by $P(w|S)$ and $P(w|L)$ respectively.

After training, the calculated probabilities are utilized to determine the probability that an email with a particular set of keywords in it belongs to either category. The probability that an email including a word *w* is spam is:

$$P(S|w) = \frac{P(S,w)}{P(w)} \quad \text{.....1}$$

$$= \frac{P(w|S)P(S)}{P(w)} \quad \text{.....2}$$

$$= \frac{P(w|S)P(S)}{P(w|S)P(S)+P(w|L)P(L)} \quad \text{.....3}$$

Then, the probability that an email including a set of keywords *W* is spam is:

$$P(S|W) = \frac{P(S,W)}{P(W)} \quad \text{.....4}$$

$$= \frac{\prod_i P(w_i|S)P(S)}{\prod_i P(w_i|S)P(S)+\prod_i P(w_i|L)P(L)} \quad \text{.....5}$$

The Bayesian filter sets a threshold, denoted by *T*. If an email's parsed keywords are *W* and $P((S|W)) \geq T$, then it is spam. Otherwise, it is considered as legitimate.

B. Identity-Based Approaches

The basic identity-based spam filtering approaches are blacklist and white list, which check the email senders for spam detection. A blacklist is a list of senders whose emails are blocked from getting through to the recipients. A whitelist is just the exact opposite. While a blacklist specifies who is to be kept out allowing all others to pass, a white list only allows those who are already on the list to get through. Since spammers almost always spoof the "From:" field of spam messages, blacklists usually keep IP addresses rather than email addresses. For incoming messages from senders not on the lists, content-based filters may be applied so that the two approaches can complement each other[5]. White lists and blacklists both maintain a list of addresses of people whose emails should not and should be blocked by the spam filter, respectively. One server side solution records the number and frequency of the same email sent to multiple destinations from specific IP addresses. If the number and frequency exceed thresholds, the node with the specific IP address is blocked [1][5].

III. PROPOSED METHOD

Radial Basis Function

The main goal of this paper is to develop improved method of spam filtering using RBF neural networks. Which is called as (RSOAP) RBF network based personalised spam filter It focuses on the impact of the widths of the Gaussian functions on accuracy of estimates generated by RBFN network.The basic model of multi-layer neural networks utilizes either sigmoid or threshold neurons. These threshold logic units (TLUs) generate a signal, or "fire," when the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

sum of their input signals exceeds a certain level. Sigmoid units provide a more graceful degradation of output signals, and because the derivative of this function can be taken the back propagation method discussed above, which relies on this value, can be implemented for training purposes[10][11].

- **RBF Architecture :**

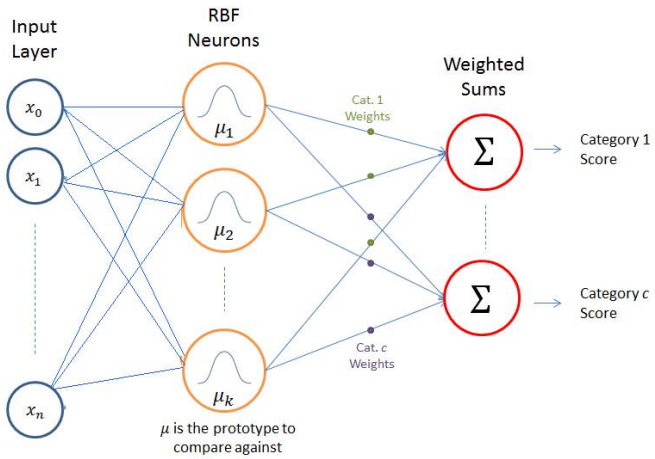


Fig 1. RBF Architecture

Figure 1 shows RBF has three layer architecture. RBF neural networks operate on a somewhat different principle. Instead of having threshold units with a single value against which to compare accumulated sums of input signals, each RBF neuron has a set of values called a "reference vector" for comparison with an input set of the same cardinality.

RBF Network have 5 parameters for optimization:

w- weights between the hidden layer and the output layer.

Φ(phi)- activation function.

μ- center of activation functions.

The distribution of center of activation functions.

M-The number of hidden neurons.

The hidden and output layers in RBF networks operate differently as compare to MLP, and the corresponding "weights" have very different meanings and properties. It is therefore appropriate to use different learning algorithms for them. The input to hidden "weights" (i.e., basis function parameters {μij ,σ j}) can be trained (or set) using any one of several possible unsupervised learning techniques. Then, the input to hidden "weights" kept fixed while the hidden to output weights are learned. The next stage in training only involves a single layer of weights {wjk} and linear output activation functions, and we have already seen how the necessary weights can be found very quickly and easily using simple matrix pseudo inversion, as in Single Layer Regression Networks or Extreme Learning Machines. One major advantage of RBF networks is the possibility of determining suitable hidden unit/basis function parameters without having to perform a full non-linear optimization of the whole network [11]. We shall now look at three ways of doing this:

1. Fixed centres selected at random
2. Clustering based approaches
3. Orthogonal Least Squares

These are all unsupervised techniques, which will be particularly useful in situations where labelled data is in short supply, but there is plenty of unlabelled data (i.e. inputs without output targets). Later we shall look at how one might try to get better results by performing a full supervised non-linear optimization of the network instead[11][12][13].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

- **Training Algorithm:**

RBF training has given in three steps[12][14]:

Step1: Choose the center randomly from the training set.

The simplest and quickest approach for setting the RBF parameters is to have their centres fixed at M points selected at random from the N data points, and to set all their widths to be equal and fixed at an appropriate size for the distribution of data points. Specifically, we can use normalised RBFs centred at {μj} defined by

$$\varphi_j(X) = exp\left[-\frac{\|X-\mu_j\|^2}{2\sigma_j{}^2}\right] \qquad \text{where } \{\mu_j\} \subset \{X^p\} \quad \ldots 6$$

Step 2 : Spreads: are chosen by normalization:

Sigma σ (spread parameter)is the standard deviation. The σj are all related in the same way to the maximum or average distance between the chosen centres μj. Common choices are

$$\sigma = \frac{\text{Maximum distance between any 2 centers}}{\sqrt{\text{number of centers}}} = \frac{d_{max}}{\sqrt{2M}}$$

which ensure that the individual RBFs are neither too wide, nor too narrow, for the given training data. For large training sets, this approach gives reasonable results.

Step 3 :Weights: are computed by means of the pseudo-inverse method.

For an example consider the output of the network.Since the hidden unit activations φ j(x,μ j,σ j) are fixed while the output weights {wjk} are determined, we essentially only have to find the weights that optimize a single layer linear network. As with MLPs, we can define a sum-squared output error measure.

$$y(x_i) = w_1\varphi_1(\| x_i - t_1 \|) + \ldots + w_{m1}\varphi_{m1}(\| x_i - t_{m1} \|) \qquad \ldots.7$$

and here the outputs are a simple linear combination of the hidden unit activations, i.e.

$$y_k(X^p) = \sum_{j=0}^{M} w_{kj}\varphi_j(X^p) \qquad \ldots 8$$

The equations for the weights are most conveniently written in matrix form by defining matrices with components. Let,

$$\Phi = \begin{bmatrix} \varphi_1(\| x_1 - t_1 \|) & \ldots & \varphi_{m1}(\| x_N - t_{m1} \|) \\ & \ldots & \\ \varphi_1(\| x_N - t_1 \|) & \ldots & \varphi_{m1}(\| x_N - t_{m1} \|) \end{bmatrix} \qquad \ldots 9$$

Then we can write

$$\Phi \begin{bmatrix} w_1 \\ \ldots \\ w_{m1} \end{bmatrix} = \begin{bmatrix} d_1 \\ \ldots \\ d_N \end{bmatrix} \qquad \ldots.10$$

If Φ^+ is the pseudo-inverse of the matrix Φ we obtain the weights using the following formula

$$[w_1 \ldots w_{m1}]^T = \Phi^+ [d_1 \ldots d_N]^T \qquad \ldots 11$$

and can be seen to have the property Φ†Φ = I. Thus the network weights can be computed by fast linear matrix inversion techniques. actually, normally it uses Singular Value Decomposition (SVD) techniques that can avoid problems due to possible ill-conditioning of Φ, i.e. ΦTΦ being singular or near singular.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

below are main steps of proposed algorithm of RFB Spam filtering:
Algorithm : RBF Spam Filter
Input
I. Test Email Features
II. Training Data Features
Output:
I. Either Legitimate Email
II. Or Spam Email
Step 1: The RBF Mapping
Step 2: The RBF Network Architecture
Step 3: Computational Power of RBF Networks
Step 4: Training an RBF Network
Step 5: Unsupervised Optimization of the Basis Functions
Step 6: Computing the Output Weights
Step 7: Supervised RBF Network Training
Step 8: STOP

Following Figure 2 shows the flowchart of Spam filtering using RBF Network. FRBF Network classifies incoming mail form dataset into Legal mail(HAM) or SPAM.

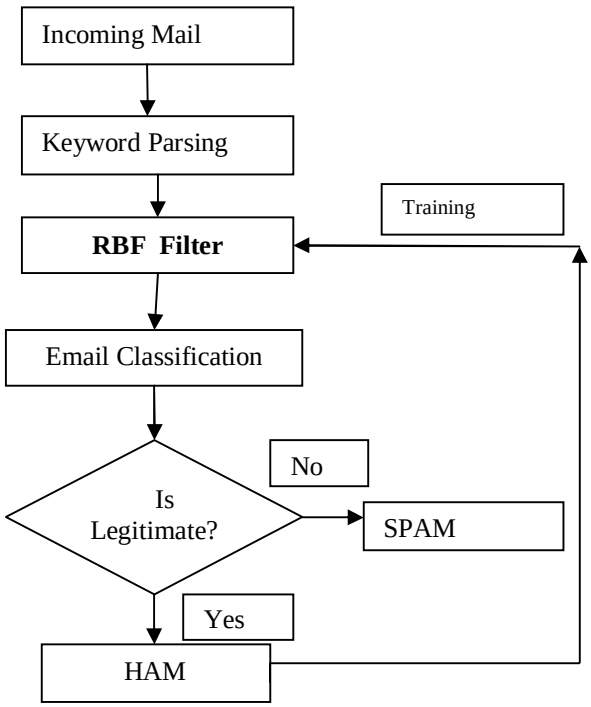


Fig 2. Flowchart of Spam Filtering Using RBF Neural network

IV. EXPERIMENTAL RESULTS

In this experiment, it has evaluated the performance of Bayesian filter compared to RSOAP (RBF network based spam filter). For that purpose it has used two different email spam datasets which are taken from different available and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

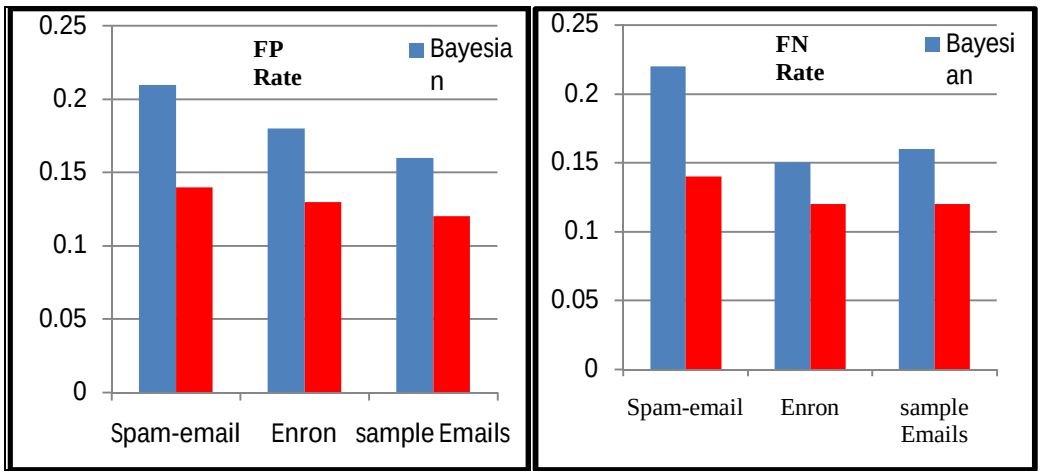
Vol. 7, Issue 6, June 2018

widely used sites. And it also tests dataset with random emails. To evaluate the effectiveness of existing Bayesian algorithm and this proposed RBF neural network algorithm uses standard evaluation measures i. e. Accuracy, Precision, Recall and F-measure.

Dataset	Method	FP Rate	FN Rate	Accuracy %	Recall %	Precision %	F-Measure %
Spam-emails	Bayesian	0.21	0.22	78	84	78	80
	RBF	0.14	0.14	86	90	86	87
Enron-Spam	Bayesian	0.18	0.15	79	86	81	86
	RBF	0.13	0.12	87	90	86	88
Testing Sample	Bayesian	0.16	0.16	82	88	84	84
	RBF	0.12	0.12	87	91	88	89

Table 1. FP Rate, FN Rate, Precision ,Recall and F-measure

Table 1 notifies FP-Rate, FN-Rate, Accuracy, Recall, precision and F-Measure value graph for all three dataset. Following figures shows the results in graphs Figure 3(a) FP Rate graph. (b) FN Rate graph. (c)Accuracy graph (d) Recall(e)Precision graph (f) F-Measure graph. It shows values for all three dataset.



(a) (b)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

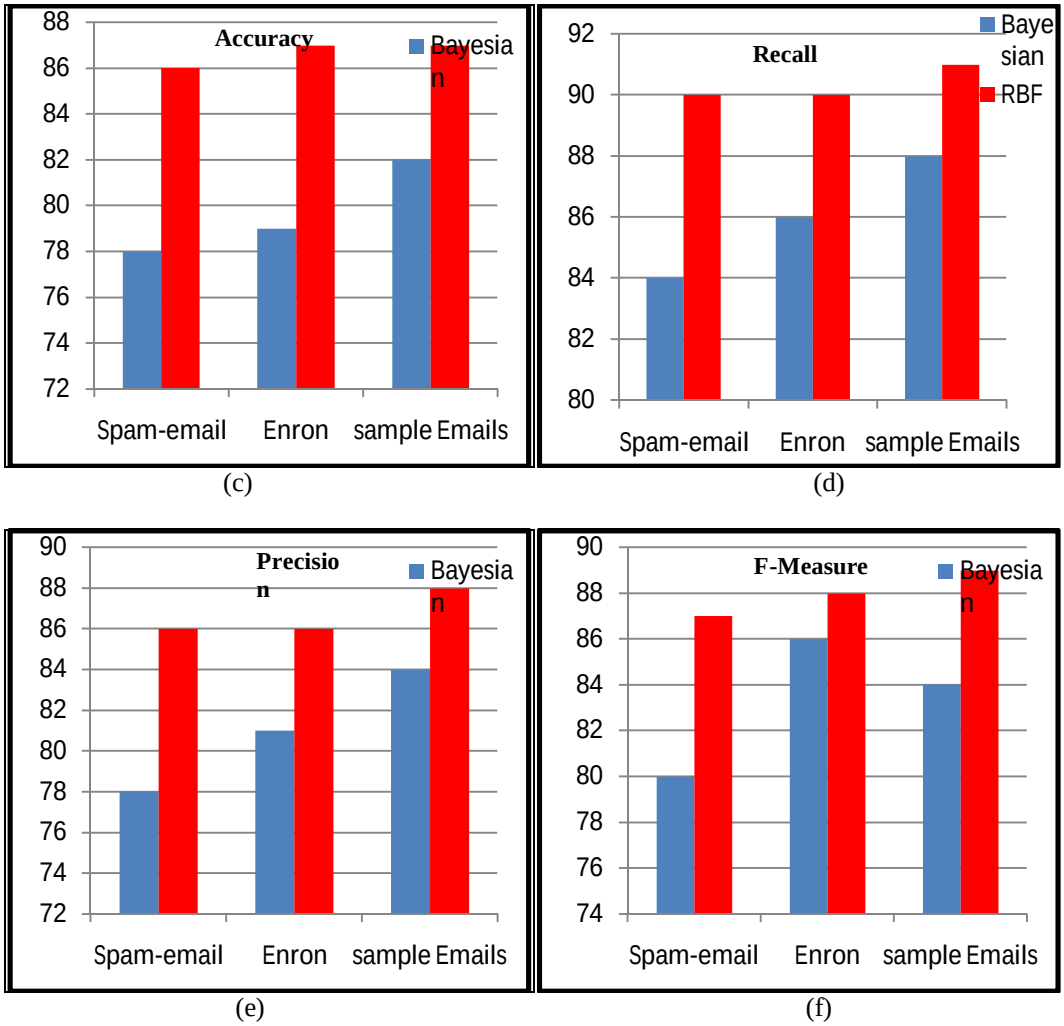


Fig 3 (a) FP Rate graph. (b)FN Rate graph. (c)Accuracy graph (d) Recall graph (e)Precision graph (f) F-Measure graph. It shows values for all three dataset.

V. CONCLUSION

Internet Spam is unsolicited email messages sent over internet to annoy user. In this paper different techniques Bayesian and RSOAP have given for spam filtering, although no single technology can achieve one hundred percent spam detection with zero false positives. By comparing existing Bayesian spam filter and proposed RBF network spam filter (RSOAP) it can say that RSOAP performs classification by measuring the input's similarity to examples from the training set. RSOAP gives less false negatives and Accuracy better as compare to Bayesian filter.

REFERENCES

[1] Haiying Shen, Senior Member, IEEE, and Ze Li, Student Member, IEEE, "*Leveraging Social Networks for Effective Spam Filtering*", IEEE TRANSACTIONS ON COMPUTERS, VOL. **63**, NO. **11**, pp. **2743-2759**, NOVEMBER **2014**.
[2] GFI Software, "*Why Bayesian Filtering Is the Most Effective Antispam Technology*,"

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

http://www.gfi.com/whitepapers/whybayesian-filtering.pdf, **2011**.

[3] M. Uemura and T. Tabata, "*Design, and Evaluation of a Bayesian-Filter-Based Image Spam Filtering Method*", Proc. Int'l Conf.Information Security and Assurance (ISA), pp. **46–51**, **2008**.

[4] J.S. Kong, P.O. Boykin, B.A. Rezaei, N. Sarshar, and V.P.Roychowdhury, "*Let Your CyberAlter Ego Share Information and Manage Spam*", Univ. of California, Los Angeles, CA, technicalreport, **2005**.

[5] H. Lam and D. Yeung, "*A Learning Approach to Spam Detection Based on Social Networks*", Proc. Conf. Email and Anti-Spam (CEAS),pp. **1–10**, **2007**.

[6] https://en.wikipedia.org/wiki/Email_spam

[7] Amit Kumar Sharma, Sudesh kumar Prajapat, Mohemmed Aslem, "*A Comparative Study between Naïve Bayes and Neural Network (MLP) Classifier for Spam Email Detection*", IJCA (**0975 – 8887**) National Seminar on Recent Advances in Wireless Networks and Communications, NWNC, pp.**12-16**, **2014**.

[8] Peiman Mamani Barnaghi, Vahid Alizadeh Sahzabi and Azuraliza Abu Bakar, "*A Comparative Study for Various Methods of Classification*", *(ICICN)*IPCSIT ,vol. **27**, pp. **62-66**, **2012**.

[9] H. Lam and D. Yeung, "*A Learning Approach to Spam DetectionBased on Social Networks*", Proc. Conf. Email and Anti-Spam (CEAS),pp. **1–10**, **2007**.

[10] Ali Idri1, Abdelali Zakrani1 and Azeddine Zahi2, "*Design of Radial Basis Function Neural Networks for Software*", IJCSI International Journal of Computer Science Issues, Vol. **7**, Issue **4**, No **3**, July **2010**.

[11] John A. Bullinaria, "*Radial Basis Function Networks: Algorithms*" Neural Computation : Lecture **14**, **2015**.

[12] Tuba Kurban and Erkan Beşdok, "*A Comparison of RBF Neural Network Training Algorithms for Inertial Sensor Based Terrain Classification*", Sensors, **6312-6329**; doi:**10.3390/s90806312**, **2009**, **9**.

[13] David P. Aguilar, "*A radial basis neural network for the analysis of transportation data* ", University of South Florida, **2004**.

[14] Foram S. Panchal, Mahesh Panchal, "*Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network*", IJCSMC, Vol. **3**, Issue. **11**, pg.**455 – 464**, November **2014**.