Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

SMS Spam Detection Using Neural Network Classifier

Anchal
Mtech,Cse,Rimt-Iet,Mandigobindgarh
India.

Abhilash Sharma
Cse,Rimt-Iet,Mandigobindgarh
India.

Abstract -SMS (Short Message Service) is a popular and quick service for the communication. The problem occurs when the user does not want to receive a particular text or text from particular type of IDS i.e is called spam . To prevent such kind of message, text classification methods have been proposed. Most of the current methods of spam detection involve using keyword filters to separate spam and regular messages. This paper focuses on the text classification methods like tree architecture, ICA algorithm and Neural Network algorithm for the text classification to prevent the user from unwanted spam messages. Although spam filtering techniques are now available on the market today, no one can deny that these solutions cannot guarantee 100% effectiveness at eliminating the problems of spam because a variety of these filters have weaknesses and strengths. The intent of this paper presents an alternative solution using a neural network classifier on a corpus of sms received by the researchers who conducted this investigation. The dataset for our system used descriptive attributes of words, symbols and sms that are commonly used by users to correctly identify spam received in inboxes.

Keywords*— *SMS, SPAM, ICA, Neural, Training, Testing.

I. INTRODUCTION

SMS is a part of our daily life. People often SMS each other to communicate. SMS can become a problem also if the user does not want to receive a SMS. Promotional companies send bulk SMS to the users which becomes a headache for the user. To identify a sms to be spam certain criteria must be decided. One of the major factors in sms spam detection is the textual analysis of the sms. E messages have become popular means for personal and business communication due to its fast and free availability as well as low or free cost. But several people and companies misuse this facility to distribute unsolicited bulk messages that are commonly called as spam sms.

Spam frustrates, confuse and annoy sms users by wasting valuable resources and time. Spam even provides ways for phishing attacks and distributing harmful content such as viruses, Trojan horses, worms and other malicious code. Several technical solutions like commercial and open-source products have been used to alleviate the effect of this issue.

Spam filtering can be of two types:

- Non-machine learning based
- Machine learning based

Short Message Service (SMS):-SMS is a communication service standardized in the GSM mobile communication systems; it can be sent and received simultaneously with GSM voice, text and image. This is possible because whereas voice, text and image take over a dedicated radio channel for the duration of the call, short messages travel over and above the radio channel using the signaling path [3]. Using communications protocols such as Short Message Peer-to-Peer (SMPP).It allows the interchange of short text messages between mobile telephone devices .

II. OVERVIEW OF TEXT CLASSIFICATION IN SMS

Text classification is a supervised learning process. In this process a task is assign on text data or document for classify this text according to predefined categories or classes according to their contents. For a long time it is very classical problem in information access field, recently this field is attracted due to over loaded amount of text document available in digital form. Some systems are based on text classification like routing, data access, classification, and filtering... So to access information from huge amount of documents is very difficult and more time consuming. Organizations that need to access information from huge amount need a technique to solve this difficulty and more work in less time. Data is automatically classified according categories of their contents. There are many algorithm are available to deal with automatic text classification [1].

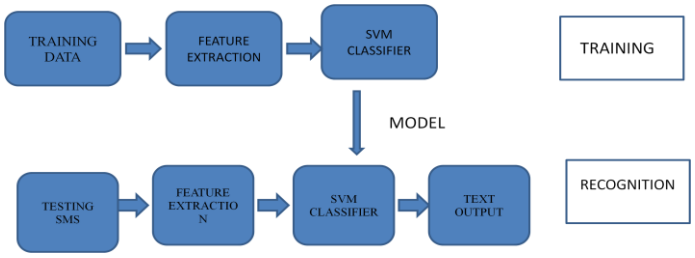

Figure 1: Spam Text Recognition system

There are several algorithms for the identification of the data text analysis:-

A. ICA(Increment Component Analysis*):*

They employ a generalized suffix-tree that can be updated efficiently when the source changes [4]. The amount of effort required for the update only depends on the size of the change, not the size of the code base... Since generalized suffix-trees are not easily distributed across different machines and the memory requirements represent the bottleneck with respect to scalability. Consequently the improvement in incremental detection comes at the cost of substantially reduced scalability.

B. Neural Logistics for text classification

The neural networks are non-linear statistical data modeling tools that are inspired by the functionality of the human brain using a set of interconnected nodes [6] networks are widely applied in classification and clustering, and its advantages are as follows. First, it is adaptive; second, it can generate robust models; and third, the classification process can be modified if new training weights are set. The neural network fraud classification model employing endogenous financial data created from the learned behavior pattern can be applied to a test sample. The neural networks can be used to predict the occurrence of corporate fraud at the management level [7].

A neural network (NN) is a feed-forward, artificial neural network that has more than one layer of hidden units [7] between its inputs and its outputs. Each hidden unit, j, typically uses the logistic

Assume that we have collected a sample of the sphere (or prewhitened) random vector function1 to map its total input from the layer below, xj , to the scalar state, yj that it sends to the layer above.

$$y_i = logistic(x_j) = 1/1 + e^{-x_j}, x_j = b_j + \sum y_i w_{ij} \quad (1)$$

where b_j is the bias of unit. j, i is an index over units in the layer below, and w_{ij} is a the weight on a connection to unit j from unit i in the layer below. For multiclass classification, output unit j converts its total input, x_j, into a class probability, p_j.

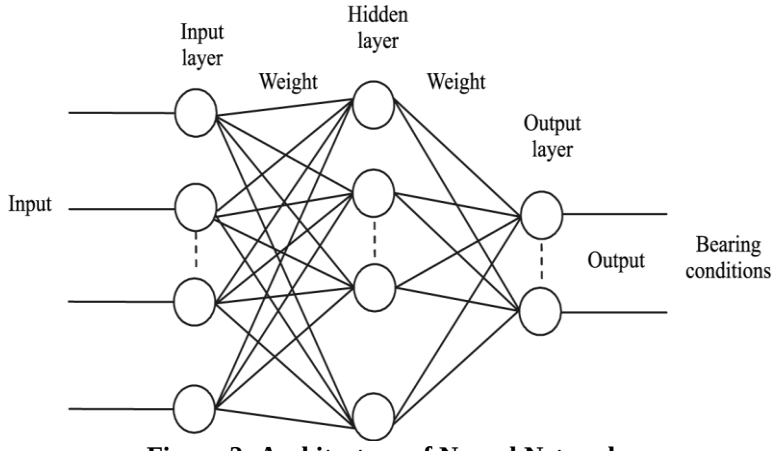

Figure 2: Architecture of Neural Network

III. WORKING OF FAST ICA ALGORITHM

FastICA is an efficient and popular algorithm for independent component analysis invented by Aapo Hyvärinen at Helsinki University of Technology. The algorithm is based on a fixed-point iteration scheme maximizing non-Gaussianity as a measure of statistical independence. It can also be derived as an approximate Newton iteration.

Fast ICA for One Independent Component

x, which is in case of blind source separation, is a collection of linear mixture of independent source signals. The basic method of Fast ICA algorithm is as follows:

- Take a random initial Vector w (0) and divide it by its norm.Let k=1.
- Let $w(k)=E\{Z[Z^T w(k-1)]^3\}-3w(k-1)$
- Divide w(k) by its norm
- If $|w^t(k)w(k-1)|$ is not close enough to 1, let k=k+1,go back to step 2.otherwise the algorithm is convergent and outputs w(k).

The final vector w(k) given by the algorithm equals one of the columns of the (orthogonal) demixing matrix B. In case of blind source separation, this means that w(k) separates one of the non-Gaussian source signals in the sense that $w(k)^T x(t)$, t = 1,2,..... equals one of the source signals.

Fast ICA for Several Independent Components

To estimate n independent components, run these algorithm n times. To ensure that we estimate each time a different independent component, we only need to add a simple orthogonalizing projection inside the loop. The column of the demixing matrix B is orthonormal because of the sphering. Thus we can estimate the independent components one by

one by projecting the current solution w(k) on the space orthogonal to the columns of the demixing matrix B previously found. Define the matrix B as the matrix whose columns are the previously found column s of B
Then adding the projection operation in the beginning of step 3.

$$\boldsymbol{w = w - B\bar{B}^{T} \times w}$$

Divide w(k) by its norm
Also the initial random vector should be projected this way before starting the iterations. To prevent estimation error in from deteriorating the estimate w(k), this projection step can be omitted after the first few iterations: once the solution w(k) has entered the basin of attraction of one of the fixed points, it will stay there and converge to that fixed point.In addition to the hierarchical (or sequential) ortho-gonalization described above, any other method of orthogonalizing the weight vectors could also be used. I n some applications, asymmetric orthogonalization might be useful. This means that the fixed point step is first performed for all the n weight vectors, and then the matrix W(k) = (w1(k),.....wn(k)) of the weight vector is orthogonalized, e.g., using the well known formula:

$$w(k) = w(k)(w(k)^{T}w(k))^{1/2}$$

Where ($W(k)^{T}W(k))^{1/2}$ is obtained from the eigenvalue decomposition of $W(k)^{T}W(k) = EDE^{T}$ as ($W(k)^{T}W(k))^{1/2} = E^{1/2}E^{T}$

IV. METHODOLOGY OF SPAM DETECTION

- First to design a spam detection system.
- To select a spam file either it is a text file or it is excel file
- To select a file on the basis of spam detection.
- Filter stemming words only from spam detection.

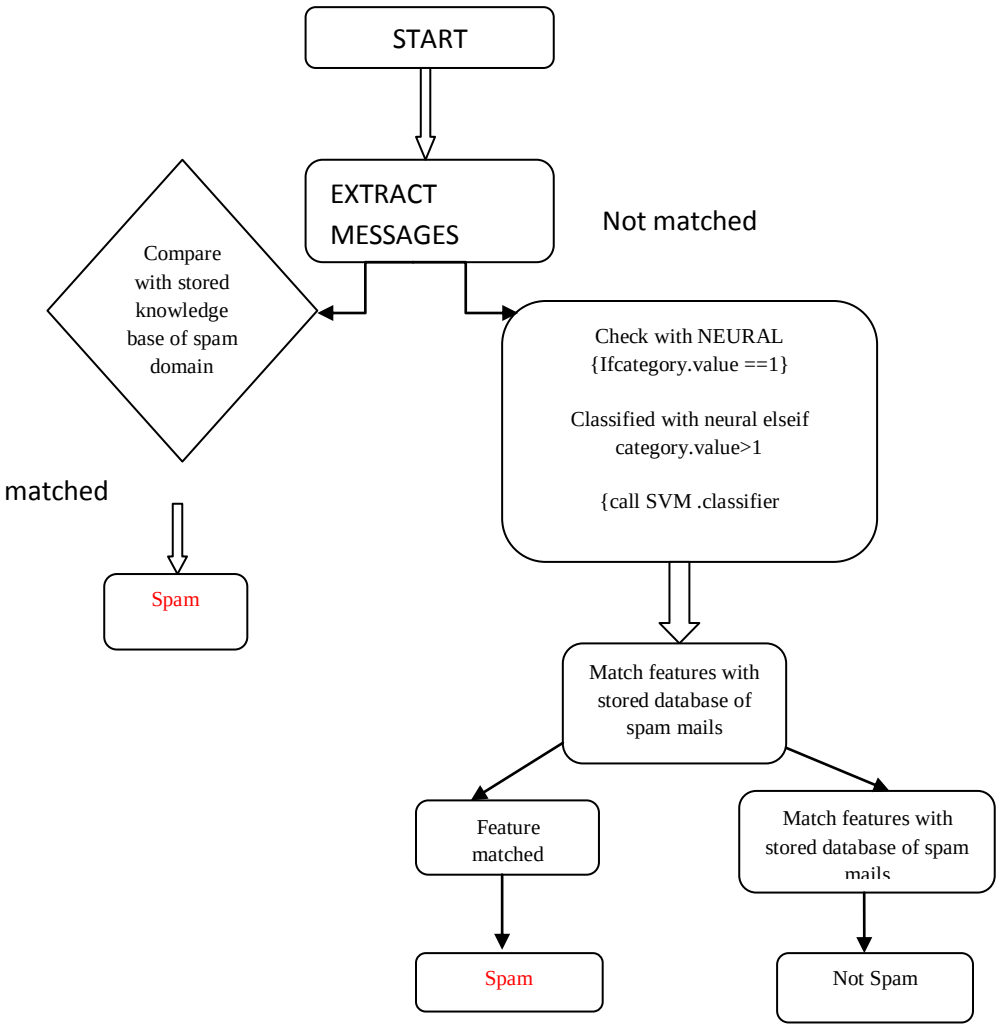

Figure 3: Flowdiagram Of Spam Detection System

Features to be extracted are on the basis of both Character-Based, Word-based, and Vocabulary features :-

1. Total no of characters (C) .
2. Total no of alpha chars / C Ratio of alpha chars.

3. Total no of digit chars / C , Total no of whitespace chars/C
4. Frequency of each letter / C (36 letters of the keyboard – A-Z, 0-9)
5. Frequency of special chars (10 chars: *, _ ,+,=,%,$,@, -, \,/)
6. Total no of words (M) .
7. Total no of short words/M Two letters or less.
8. Total no of chars in words/C
9. Average word length .
10. Avg. sentence length in chars .
11. Avg. sentence length in words
12. Word length freq. distribution/M Ratio of words of length n, n between 1 and 15
13. Type Token Ratio No. Of unique Words/ M
14. Frequency of punctuation 18 punctuation chars:. ‹; ? ! : ( ) – " « » < > [ ] { }

V. CONCLUSION

In this paper we have reviewed the text classifiers such as FastICA for feature extraction and neural networks. A neural network system is useful and accurate tool for classifying spam messages but to enhance precision performance, supervision is needed.It requires fewer input features to achieve the same results produced by other classifiers.The overall aim of this paper to explore the idea of text classification methods like fast ICA and neural networks for pattern matching, feature extraction to increase accuracy of the system. Furthermore our aim will be using classifiers of neural networks to increase the efficiency of spam detection system.

VI. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my guide Mr. Abhilash Sharma for his valuable guidance and advice.Also I would like to thank to all the people who have given their heart welling support in making this completion a magnificent experience.

REFRENCES

[1]. Qian Xu and Evan Wei Xiang, Baidu "SMS Spam DetectionUsing Noncontent Features" 1541-1672/12/$31.00 © 2012 IEEE INTELLIGENT SYSTEMS Published by the IEEE Computer Society

[2] Zhenhai Duan, Senior Member, IEEE, Peng Chen, Fernando Sanchez,Yingfei Dong, Member, IEEE, Mary Stephenson, and James Michael Barker , "Detecting Spam Zombies by Monitoring Outgoing Messages", IEEE APRIL 2012

[3] H. A. Basit, D. C. Rajapakse, and S. Jarzabek. "A study of clones in the STL and some general implications". In Proc. of the Int'l Conf. on Software Engineering, "pages 451{459, 2005.

[4] I. D. Baxter, A. Yahin, L. M. de Moura, M. Sant'Anna, and L. Bier. "Clone detection using abstract syntax trees". In Proc. of the Int'l Conf. on Software Maintenance " pages 368{377, 1998.

[5] K. Beck. "Extreme Programming explained, embrace change" Addison- Wesley, 2000.

[6] Hang Dai and Jingshi He Dongguan " China Research Journal of Applied Sciences, Engineering and Technology"6(5): 895-899, 2013 ISSN: 2040-7459; e- ISSN: 2040-7467 2013.

[7] T. T. Nguyen, H. A. Nguyen, J. M. Al-Kofahi, N. H. Pham, and T. N. Nguyen, "Scalable and incremental clone detection for evolving software," ICSM'09, 2009.

[8] Ghosh, S., & Reilly, D. L. (1994). "Credit card fraud detection with a neural- network", 27th Annual Hawaii International, Conference on System Science 3 (1994) 621–630.

[9] Beasley, M. (1996)." An empirical analysis of the relation between board of director composition and financial statement fraud" The Accounting Review",71(4), 443–466.

[10] J. H. Johnson, "Identifying redundancy in source code using fingerprints,"in Proc. of CASCON '93, 1993, pp. 171–183.

[11] M. Fisher and G. Rothermel, "The EUSES spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms," ACM SIGSOFT Software Engineering Notes, vol. 30, no. 4, pp. 1–5,2005.

[12] I. D. Baxter, A. Yahin, L. M. de Moura, M. Sant'Anna, and L. Bier, "Clone detection using abstract syntax trees," in Proc. of ICSM '98, 1998, pp. 368–377.

[13] R. Komondoor and S. Horwitz, "Using slicing to identify duplication in source code," in Proc. of SAS '01, 2001, pp. 40–56.

[14] G.D.K.Kishore1, Maddali Sravanthi Automated Anomaly and Root Cause Detection in Distributed Systems.International journal of engineering trends and technology-Volume3Issue1-2012.s

[15] David Ndumiyana, Munyaradzi Magomelo and Lucy Sakala,"Spam detection using a neural network classifer"in ISSN 2315-5027;Volume 2;Issue 2,April 2013.

[16] Punskis D, Laurutis R, Dirmeikis R, An Artificial Neural Nets for Spam email Recognition, Elect Elect Engine,2006; 5(69): 1392–1215.

[17] Wu,C..,Cheng,K.T.,Zhu,Q.,Wu,Y.L,"UsingVisual Features For Anti-Spam Filtering," 2005 IEEE International Conference on Image Processing (ICIP 2005),pp.509–512,2005.

[18] Toshihiro Tabata,"spam mail filtering: commentary of Bayesian filter,"The journal of information and science and technology association,Vol 56 ,no.10, pp464-468,2006.

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

[20 Information Theoretic Feature Extraction", Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July-Aug,2005.

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

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

[23] 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).

