Detection of Spams using Extended ICA & Neural Networks

Deepinderjeet Kaur[1], Amandeep Kaur[2]
[1]Research Fellow, [2]Asst. Professor
[1,2]Sri Guru Granth Sahib World University,Fatehgarh Sahib,Punjab.

Abstract* — Spams are the textual context of the system which can damage the system. The basic problem is to protect the system from such type of unwanted files. To save from system form such kind of failures we design a system which can recognize the spams and can let you know on the basis of training system. The first part consists of filling the ip address or header in the text box. In the ip.txt file we put the ip addresses of those countries or region which we want to be marked as spam and on the other hand in header.txt file we put the headers of all our contacts. In the second part we detect the spam as we compare the content of the given file with the spam.txt file. In the spam.txt file we put the spam words. For detection purposes, we used ICA++ algorithm and for matching purpose, we used Neural Networks. If the 70% of the data of the given file matches with the spam words then it is declared as spam file and at the end there are comparison between PCA & ICA++, first on the basis of max error rate second on the basis of processing time third on the basis of accuracy.

***Keywords*— Spam, Detection of Spams, ICA++, PCA, Neural Network.**

I. INTRODUCTION

At one purpose or another – just like the majority of laptop users – you have got received emails that promise business deals price various pounds, that try and sell product to enhance your look or that try and win over that its price investment your cash during a specific company or stock. Handling spam (unsolicited email that's not targeted at specific individuals), is one downside that every one email users share in common. Analysis shows that between sixty fifth and ninetieth of all email received is taken into account spam. On a personal user basis, spam is annoying; it's a waste of your time and infrequently contains spyware, malware and even porno. On a company-wide basis, constant threats apply but there's additionally the money price to manage spam that has to be taken into thought.

1.1 The evolution of spam

Until a minute past, spam was the domain of text- or html-based emails. For anonymous delivery, these messages historically relied on abusing open SMTP relays. Once open SMTP relays became less common, spammers switched to proxy servers, dial-up services and a lot of recently, hijacked computers. Spammers designed personalised template emails to deliver their messages and so created use of bulk mailing software system for distribution.

To block spam, email service suppliers and firms typically relied on keyword 'detection', and thespian up an inventory of keywords that ordinarily appeared in most of the spam email. This list would typically embrace keywords like 'viagra' or 'bank'. However, this methodology typically blocked real email and adding a lot of keywords merely resulted in additional false positives that successively blocked legitimate email. However spammers became smarter too, and that they self-addressed keyword interference by commutation keywords like 'vi@gra' to 'v1agra'.

Another try at interference spam includes creating use of blacklists that contain an inventory of IP addresses of notable spammers or compromised hosts. However, these lists need to be perpetually updated as a result of spammers has learnt to counteract this by quickly ever-changing the origin of spam.

1.2 The latest trends

Although spammers registered tidy success with image spam (picture, right) the anti-spam software system trade had not lost the battle and quickly came out with new counter-measures to prevent image spam. Realizing that filters had a drag with pictures, the solution was to hit spammers at supply – that's wherever the e-mail originated from. This new approach had an on the spot positive result and significantly weakened the effectiveness of image spam and gave back to email users some management over their mailbox.

As with each cat-and-mouse game, spammers had to reply and in June 2007, they came up with a replacement technique that's not solely ingenious however even a lot of problematic than image spam. Rather than embedding the image inside the e-mail itself, they 'repackaged' it inside Associate in Nursing attachment victimisation one in every of the foremost common file formats in use these days – a PDF file. Email users 'expect' spam to be a picture or text inside the body of the e-mail not an attachment.

Since most businesses these days transfer documents victimisation the PDF format, email users can got to check every PDF document otherwise they risk losing necessary documentation. With most anti-spam software system merchandise on the market intermeshed towards filtering the e-mail itself and not attachments, spam features a longer shelf-life inside a network.

Associate in Nursing attachment that's a PDF file has bigger credibleness in Associate in Nursing email so creating social engineering attacks a lot of easier. the flexibility to send giant PDF files may end in one spam attack inflicting large bottlenecks on a company's email server, reducing the standard and quantity of information measure on the market. By causation PDF attachments, spammers may also resort to phishing by attaching purportedly authentic documents from a bank or service supplier.

The use of PDF spam was transitory as anti-spam software system vendors quickly came out with updates and filters that analyzed the body of each PDF file. to not be defeated, spammers took but a month to come back out with a replacement option: Microsoft surpass files for push-and-dump scams. This move was clever for reasons almost like those higher than for PDFs:

- Email users 'expect' spam to be a picture or text inside the body of the e-mail Associate in Nursingd not an attachment.

- Excel is another extraordinarily common file-type in use and users square measure terribly acquainted with this format.

Since several businesses use Microsoft surpasses for spreadsheets, Spambases and then on, email users can get to check every document otherwise they risk losing necessary documentation. With most anti-spam software system merchandise on the market intermeshed towards filtering the e-mail itself and not attachments, 'Excel' spam features a longer shelf-life inside a network. Taking the sport to a replacement level, in early August 2007, spammers started pressure their text-based and Excel-based spam documents victimisation the nada file format. this can be effective for 2 main reasons:

- Companies that don't use anti-virus software system on their network may be straightforward targets for this sort of spam.
- Users World Health Organization might not bear in mind of security problems encompassing attachments square measure vulnerable to gap these nada files.

With spammers and hackers thriving in their unholy alliance, the chance of malicious files being packaged with pump-and-dump spam is only too real.

II. Review of Related Work

Mrs. Bharati M. Ramageri [1] provides that Spam Detection has importance concerning finding the patterns, statement, discovery of data etc., in numerous business domains. Spam Detection techniques and algorithms like classification, bunch etc., helps to find the patterns to come to a decision upon the long run trends in businesses to grow. Spam Detection has wide application domain nearly in each business wherever the Spam is generated that's why Spam Detection is taken into account one amongst the foremost necessary frontiers in Spambase and knowledge systems and one amongst the foremost promising knowledge domain developments in info Technology. Various algorithms and techniques like Classification, Clustering, Regression, computing, Neural Networks, Association Rules, call Trees, Genetic rule, Nearest Neighbour technique etc., area unit used for data discovery from Spambases [1].

Clustering is that the method of grouping a collection of Spam objects into multiple teams or clusters in order that the objects among the cluster have high similarity, however area unit terribly dissimilar to things in different cluster.Various bunch technhiques area unit enforced and analysed employing a bunch tool wood hen [9].

Speaker diarization analysis has been performed with several approaches and techniques. All of them area unit characterised in options extraction, speaker segmentation, and speaker bunch. Speaker diarization may be a Spam-processing technique. There aret several well-known acoustic options are used within the audio signal extraction. Mel Frequency Cepstral Coefficients (MFCCs) is that the most typical acoustic feature selection for speaker diarization. Speaker segmentation is usually performed by victimization energy-based, model-based, or measure-based segmentation. within the energy-based segmentation, the analysis relies on the acoustic energy of the audio stream. The common selection for model-based approaches is applying Hidden Andre Markoff Model (HMM) and Gaussian Mixture Models (GMMs). GMMs area unit ofttimes chosen as a result of the universal density approximators, i.e., they\'ll model associate degree capricious likelihood distribution operate over the Spam. Measure-based approaches live the distinction between 2 consecutive segments of the audio stream that is sometimes mentioned as distance between the 2 segments. Speaker bunch approaches is classified into hierarchical-based and model-based bunch. The ranked bunch is associate degree intuitive technique to cluster the audio segments. Initially, the gap between every try of Spam segments is computed, employing a user outlined distance live that assigns smaller distances to acoustically similar segments. The model-based bunch is usually performed in parallel with the segmentation. Multiple passes over the audio stream produce the optimum segmentation, cluster the Spam, alter the cluster model parameters and re-segment the audio track. This procedure is iterated till convergence or till a stopping criterion is met.

Spam text recognition aims to mechanically establish the textal state of a personality\'s being from his or her voice. it\'s supported in-depth analysis of the generation mechanism of Spam signal, extracting some options that contain textal info from the speaker's voice, and taking applicable pattern recognition strategies to spot textal states. Like typical pattern recognition systems, our Spam text recognition system contains four main modules: Spam input, feature extraction, SVM based mostly classification, and text output.

[11] Björn Schuller, Gerhard Rigoll, and Manfred Lang says that Spam text recognition is one of the latest challenges in Spam processing. Besides human facial expressions Spam has proven as one of the most promising modalities for the automatic recognition of human texts. Especially in the field of security systems a growing interest can be observed throughout the last year. Besides, the detection of lies, video games and psychiatric aid are often claimed as further scenarios for text recognition . Addressing clustering in a practical view it has to be considered that a technical approach can only rely on pragmatic decisions about kind, extent and number of texts suiting the situation. It seems reasonable to adapt and limit this number and kind of recognizable text to the requirements given within the application to ensure a robust clustering. Yet no standard exists for the clustering of texts in technical recognition. An often favored way is to distinguish between a defined set of discrete texts. However, as mentioned, no common opinion exists about their number and naming. A recent approach can be found in the MPEG4 standard, which names the six texts anger, disgust, fear, joy, sadness and surprise. The addition of a neutral state seems reasonable to realize the absence of any of these texts. This clustering is used as a basis for the comparison throughout this work also expecting further comparisons. Most approaches in nowadays Spam text recognition use global statistics of a phrase as basis. However also first efforts in recognition of instantaneous features exist. We present two working engines using both alluded alternatives by use of continuous hidden Markov models, which have evolved as a far spread standard technique in Spam processing.

[12] Indian Institute of Technology, Kanpur2LTI, School of Computer Science, Carnegie Mellon University, Pittsburgh3International Institute of Information Technology, Hyderabad. They have something different to say about this. They see the process of communication in voice processing like this:
AdaBoost algorithm is an adaptive classifier which iteratively builds a strong classifier from a weak classifier. In each iteration, the weak classifier is used to classify the Spam points of training Spam set. Initially all the Spam points are given equal weights, but after each iteration, the weight of the incorrectly classified Spam points increases so that the classifier in the next iteration focuses more on them. This results in decrease of the global error of the classifier and hence builds a stronger classifier. The Berlin Textal Spamset consists of Spam files from 10 different speakers- 5 males and 5 females. For the training purpose, we use Spam files from 3 males and 3 females to create gender independent classifiers. After extraction of the 70 features explained above, we take two texts at a time and build a binary classifier for them.

[13] International Journal of Advanced Engineering Research and Studies E-ISSN2249–8974 Support Vector Machine is simple and efficient computation of machine learning algorithm, and used in the pattern recognition and clustering issues. SVM is having the advantage that for the limited training Spam, it is having very good clustering performance. The idea behind the SVM is to transform the original input set to a high dimensional feature space by using kernel function. Therefore non linear problems can be solved by doing this transformation . Following figure shows the support vector machine with kernel function, in which input space is consisting of input samples converted into high dimensional feature space and therefore input samples become linearly separable.

III. PROPOSED METHODOLOGY

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

3.1 Proposed Model

The proposed model focuses on following objectives which are helpful for detection of spams.

a) To design a system for spam detection of spam.
b) To train the system about the spam.
c) To check the documents on the basis of stemming, and pattern matching.
d) To increase the accuracy of the spam detection.

3.2 Basic Block Design

In this proposed work, the basic purpose is to detect the spams. The basic design of the system is as shown in Fig 1. In this, text file & spam detection file, both are firstly converted into the signal using ICA++ (Independent Component Analysis) algorithm. Then matching is done using neural network. Neural Networks are able to discover rules which otherwise are difficult for humans to describe or even comprehend. They not only provide an easy way to model complex relationships between input and output, but also provide adaptability and learning ability implicitly. If the 70% of the data of the given file matches with the spam words then it is declared as spam file.

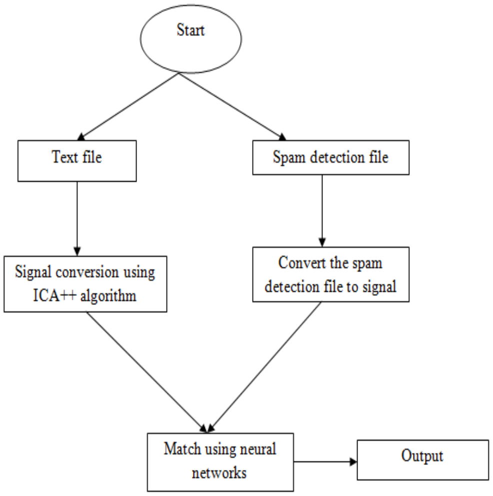
Start
Text file
Spam detection file
Signal conversion using ICA++ algorithm
Convert the spam detection file to signal
Match using neural networks
Output

Fig 1: Block Design of Proposed Approach

3.3 Algorithm Level Design

The algorithm level design of this work is shown in Fig 2. This proposed work is divided into three phases. First phase is 'to process content file', second phase is 'to process signal file' and third phase is 'matching phase'.

Phase 1: To Process Content File

STEP1: Upload Content File.

STEP2: Divide the text of content file into clusters.

STEP3: Use ICA algorithm to change data to signal conversion.

Phase 2: To Process Spam File

STEP1: Upload Spam file.

STEP2: Use ICA algorithm to change data to signals.

Phase 3: For Matching

STEP1: Proceed after signal conversion.

STEP2: Check Threshold. (Default threshold=70)

STEP3: Call Neural classifier for matching using match (exceeds threshold) Match++.

STEP4: Compute Accuracy time & then compare the results with PCA.

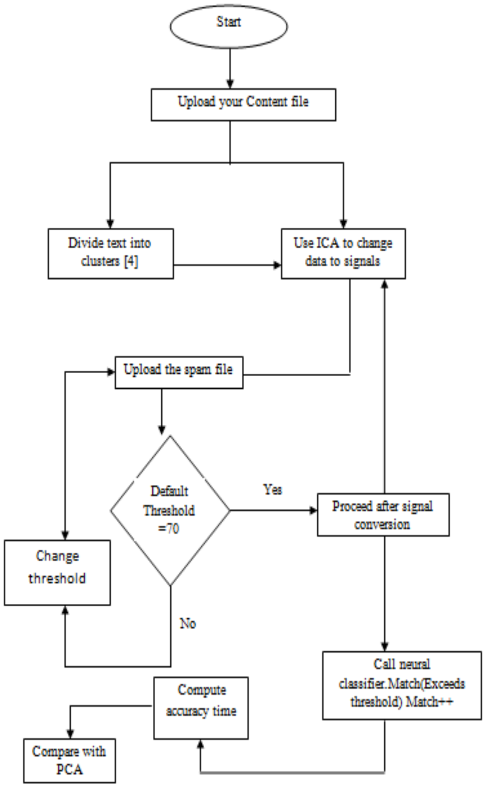
Start
Upload your Content file
Divide text into clusters [4]
Use ICA to change data to signals
Upload the spam file
Default Threshold =70
Yes
Proceed after signal conversion
Change threshold
No
Call neural classifier.Match(Exceeds threshold) Match++
Compute accuracy time
Compare with PCA

Fig 2: Algorithm Level Design

IV. RESULTS

The basic purpose of proposed work is to detect the spams. First part consists of filling the ip address or header in the text box. In the ip.txt file we put the ip addresses of those countries or region which we want to be marked as spam and on the other hand in header.txt file we put the headers of all our contacts.header.txt and ip.txt are vice versa files. Like in ip.txt file we put the ip addresses of those countries which we want to mark as spam but in header.txt file we have email addresses of our contact list. So the mails received from outside the header.txt file are marked as spam. In the second part we detect the spam by compare the content of the given file with the spam.txt file. In the spam.txt file we put the spam words. If the 70% of the data of the given file matches with the spam words then it is declared as spam file and the comparison graphs on the basis of max error rate, processing time and accuracy are shown below.

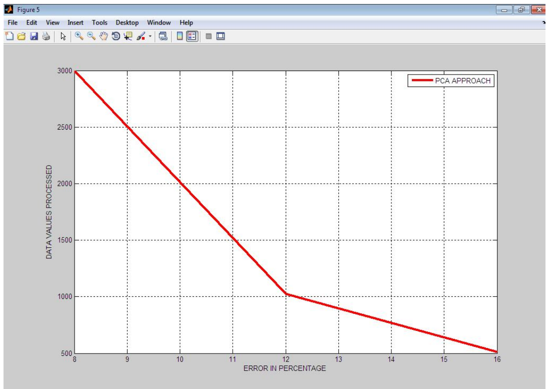

Fig 3: Max Error Rate (PCA)

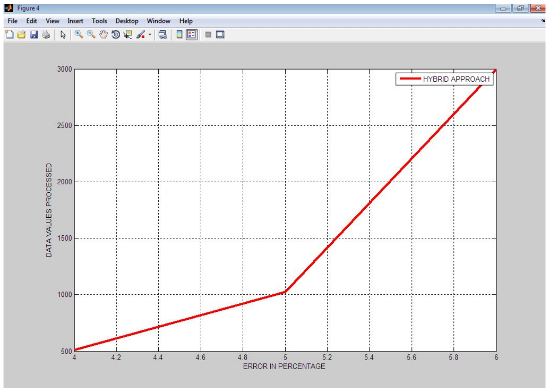

Fig 4: Max Error Rate (Hybrid)

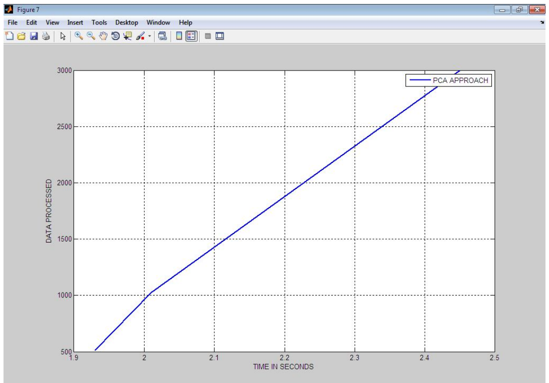

Fig5: Time Plot (PCA)

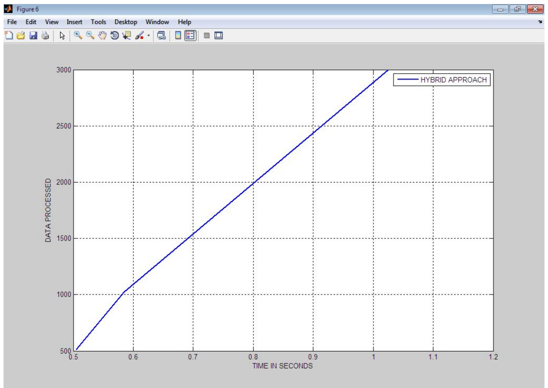

Fig 6: Time Plot (Hybrid)

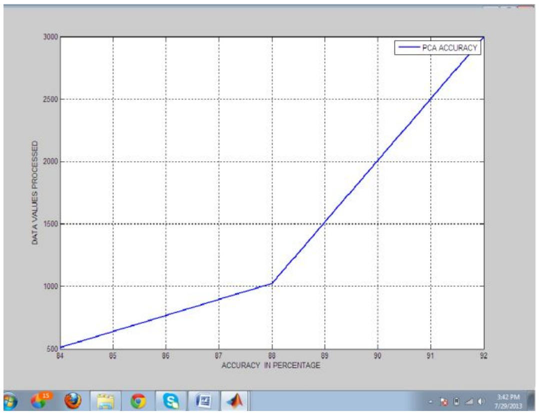

Fig 7: PCA Accuracy

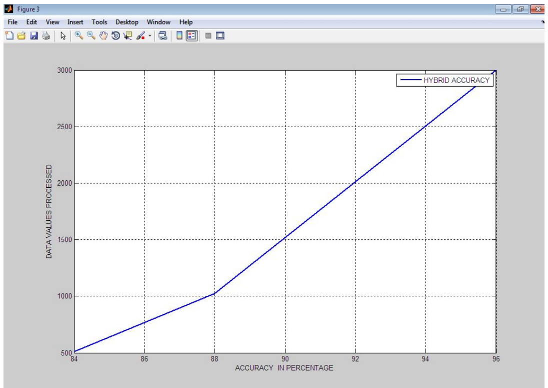

Fig 8: Hybrid Accuracy

V. CONCLUSION

The amount of junk e-mail, commonly called spam, has skyrocketed in the recent past. Traditionally, spams sent by single source mass mailers (spammers). The basic purpose of our proposed work is to detect spams using extended ICA and Neural Networks. The results show that our results are better than the previous approach where PCA is used in place of extended ICA algorithm for signal conversion.

REFERENCES

[1] Viruslist.com (2007), *Contemporary Spammer Technologies* available from:
http://www.viruslist.com/en/spam/info?chapter=153350528

[2] NetworkWorld.com (2007), *Spam Calculator* available from: http://www.networkworld.com/spam/index.jsp

[3] SecureWorks (2007), *Storm Worm DDoS Attack* available from:
http://www.secureworks.com/research/threats/view.html?threat=storm-worm

[4] The TechWeb Network (2007), *Dutch Botnet Suspects Ran 1.5 Million Machines* available from:
http://www.techweb.com/wire/security/172303160

[5] BBC News website (2007), *Criminals 'may overwhelm the web'* available from:
http://news.bbc.co.uk/2/hi/business/6298641.stm

[6] Bächer P., Holz T., Kötter M. and Wicherski G. (2007), *Know your Enemy: Tracking Botnets* available from:
http://www.honeynet.org/papers/bots/

[7] White Papers . Spam Tutorial. . VicomSoft.
http://www.spambolt.com/anti_spam_faq/ email_spam_filter.html

[8] Sivanadyan, Detecting spam using Neural Networks.
http://www.cae.wisc.edu/~ece539/project/f03/sivanadyan.pdf

[9]Patrick Pantel and Dekang Lin.‘‘SpamCop.A Spam Classification & Organization Program. Proceedings of AAAI-98, Workshop on Learning for Text Categorization.

[10] Paul Graham, at the 2003 Spam Conference http://www.paulgraham.com/better.html

[11] Lindsay I Smith. A tutorial on Principal Components Analysis
http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

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

[13] The Great Spam Archive (Online spam database) http://www.annexia.org/spam/