Text and Image based Spam Email Classification using an ANN Model- an Approach

Mr. Rahul Bansod,
CSE Department
BDCOE, Sewagram
Wardha, India
rahulbansod15@gmail.com

Prof. R. S. Mangrulkar
CSE Department
BDCOE, Sewagram
Wardha, India
rsmangrulkar@gmail.com

Prof. V. G. Bhujade
CSE Department
BDCOE, Sewagram
Wardha, India
vaishali.hardeo@gmail.com

***Abstract*-**Email has become one of the fastest and most economical forms of communication. The increase of email users have resulted in the dramatic growth of spam emails over the recent few years. The mass of e-mail that we get is rapidly increasing. People are wasting much time in filtering e-mails and organizing them into folders in order to facilitate fast retrieval. The rate of unsolicited e-mails is also growing rapidly. For today's systems the traditional way of spam detection based on signature is no more efficient. For achieving computer security now researchers are interested in the field of immune system. The function of computer security system is developed to recognize and discard spam. Image spam is kind of spam invented by the spammers where advertising details are specified in the image. Mass unsolicited electronic mail called spam has increased enormously in a short time ago and has become a serious threat not only to the Internet but also to society. In this paper a method is proposed for the classification of text and image based spam mails using Artificial Neural Network (ANN). Training and Testing will be performed on two data sets one for text based mails and another for image based mails. OCR tool is required for text extraction from image.

Keywords*-Artificial Neural Network, Spam, Spam Filter, OCR.*

*****

I. INTRODUCTION

The use of internet is an efficient form of communication that has been widely adopted by both individuals and organizations over the past decade and it continues to be on the ascent. Hence, Internet is gradually becoming an integral part of everyday life. Today, more and more people are relying on e-mail to connect them with their friends and family. Internet usage is expected to continue growing and e-mail has become a powerful tool intended for idea and information exchange.

Fig. 1 Normal images

Fig 2: Spam images

An email usage has evolved, due to its low costs, negligible time delay during transmission and security of the data being transferred. But still there are few pitfalls that spoil the well-organized consumption of email. One of them is Spam email. The extreme effects of spam emails are on the loss of individual productivity and also the financial loss of organizations. Anti-spammers, therefore, are putting forward efforts to prevent this potential threat to today's Internet. The service providers and organizations are spending billions of dollars per year in lost bandwidth which makes it an expensive problem. Because of the many spam filters which have been originated these days, to detect the advertising text in the text mail, spammers have launched a new category of spam, where the advertising text is embedded in the image. This type of mail is an image spam where the textual spam message is embedded into images.

There are several approaches which try to stop or reduce the large amount of spam arrival in the client's system. Anti-spam laws can be applied universally to decrease this bulk amount of arrival. Other techniques are based on network information and IP addresses in order to detect whether a message is spam or ham. Filtering techniques based on the contents of the spam are the most common techniques to identify whether a message is spam or not [6]. Several Machine Learning techniques such as Bayesian classifiers and Support Vector Machine [7] are also applied to the problem of spam email. This paper presents a technique for both texts based and image based spam email classification.

II. RELATED WORK

Several attempts in the literature have been suggested for solving the problem of the spam. Sarab M. Hameed1 et al. [6] has proposed an automated tool named a spam filter is using (OBP) "Optical Back Propagation" technique to identify whether a message is spam or not based on the content of the message. The OBP algorithm is designed to overcome some of the problems associated with standard "Back-Propagation". The important property of this algorithm is that it can escape from local minima during the training period with the high speed of convergence. By adjusting the error, the convergence

speed of the learning process can be improved. Thiagarajan Sivanadyan et al. [2] have proposed a text classification method, which searches the textual content of an email and also employs an algorithm to classify an email as spam or not-spam. The algorithm is capable enough to classify the occurrence of certain words and phrases in terms of how and where they appear in the email message.

James Clark et al. [3] has presented a highly flexible NN-based system for e-mail classification. It works in two phases preprocessing and classification. All unique words in the entire training corpus are identified. Feature Selectors such as information gain (IG) and variance (V) are used for feature selection. The most important words are chosen by applying feature selection. Each document is then represented by a vector that contains a normalized weighting for every word according to its importance. Weighting schemes like term frequency and inverse document frequency are implemented.

Ngo Phuong Nhung, Tu Minh Phuong [4] have proposed a method to detect spam images by using an edge-based feature vector, to represent major shape properties of the image. A vector of similarity measures are calculated using the edge-based feature. SVM considers these similarity vectors as an input for classification. The method does not use computationally expensive image processing hence it is fast.

Xiao Mang Li et al. [5] has proposed an approach on several public spam corpus and verified the effectiveness in terms of the filtering capacity and performance. OCR module is inserted into the traditional text-based filter to extract embedded text in image attachments. The extracted text is then combined with other textual contents in the email body. Eventually a text classification is applied to calculate the spam-like probability of the whole textual content.

III. THE PROBLEMS DUE TO SPAM MASSAGES

Spam email is a widespread problem on the Internet. Spam email is so cheap to send, this uninvited messages are sent to a large number of users unmethodically. When a large number of spam messages are received, it is necessary to take a long time to identify spam or non-spam email and these email messages may cause the mail server to crush.

Today, there are many types of spam mails, for example, advertisements for the purpose of earning money or selling something, and also for the purpose of spreading pranks. The spammer sends spam mails for advertising goods, services and ideas. Many times spammers can use spam mails to cheat people out of their confidential information, to deliver spiteful software, or to cause a temporary failure of a mail server. Variety of phenomena shows that the negative impact of spam messages has been seriously disrupting people's normal work and life. Spam messages brought a very bad influence on social harmony. Some criminals use spam message to spread rumors and to provoke ethnic hatred, which influences the social communism.

Spam messages are sent in mass, so the transmission time takes much network bandwidth, causing congestion, influencing the performance of the network and also the people's normal correspondence. Spammers may engage in deliberate fraud to send out their messages. Spammers often use fake names, phone numbers, addresses, and other contact information to set up "disposable" accounts at various internet service providers. They use falsified or stolen credit card numbers many times to pay for these accounts. This permit those to move quickly from one account to the next as the host ISPs discover and shut down each one. Spammers frequently seek out and make use of vulnerable third-party systems such as open mail relays and open proxy servers.

IV. PROPOSED SYSTEM

The proposed system will identify both text and image based spam. Two datasets are used in the system. One for text based spam and another for image based spam. The system can also process the running mails. The neural network will be supervised after every mail. The newly arrived text based spam will be black listed, after analyzing the advertising data in the document. The image based spam is processed firstly by extracting text from the image using OCR tool and after applying certain preprocessing over the extracted text the mail is identified as spam or ham based on the nature of text.

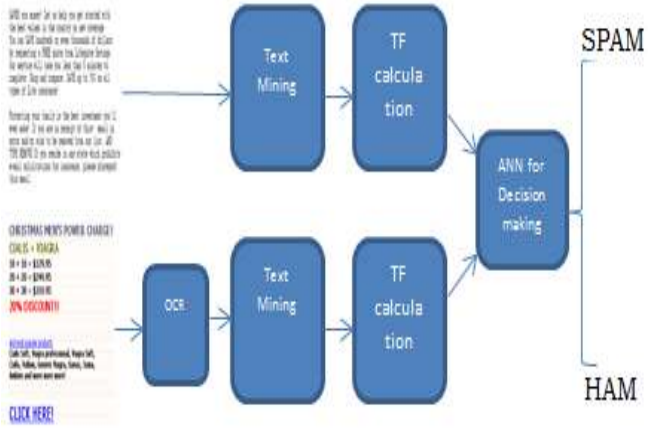

Fig. 3 Proposed System

A. *Black Listing and White listing*

A domain list is maintained of all those email addresses which are mostly sending spam mails. All those web pages and domain that are widely known for sending spam mails and are not trusted, go onto the black list. The mail is predicted spam, without any further processing if a domain matches from this list.

Black Listing: Black-listing is creating a list of domain names, when a mail comes from that specific domain, it is considered spam. No further processing is done.

White Listing: White list is a list of trusted domains and a mail from them is always ham. White listing is a method used to classify user's email addresses as legitimate ones.

B. *Words extraction from Images*

Users have an option of attaching image to their mails. The image is passed through the google's open source library Tesseract, and words are extracted from it. These words are then passing through the artificial neural network to predict the mail as spam or ham. Optical character recognition (OCR) method has been used to get editable text from printed text.

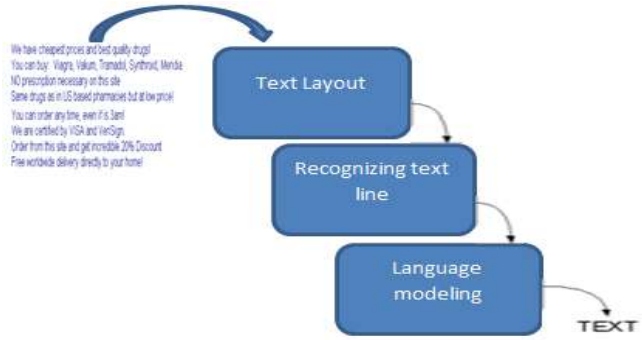

Fig. 4 Flow diagram of OCR engine

The OCR system works with three major components.

- Physical layout analysis is answerable to identify text blocks, text lines, text columns and reading order.
- Text line recognition component recognizes the text contained within each line.
- Statistical language modeling assimilates alternative recognition hypotheses with prior knowledge about language, grammar, vocabulary, and the domain of the document.

C. *Stemming Operation*

A database of all the words that occur in each mail with the frequency of the word stored in each column will be maintained. The words are converted to their root form by applying Stemming.

Some steps of this process are:

- Remove the plurals and -ed or –ing suffixes
- Deal with suffixes, -full, -ness etc.
- Take off -ant, -ence etc.

D. *Stop words removal*

Stop words are language specific functional words which carry no information. It may be of the types such as pronouns, prepositions, conjunctions. These are some of the most common words, such as *the*, *is*, *at*, *which*, *on etc*. These most frequently used Stop words in English are useless in Text mining.

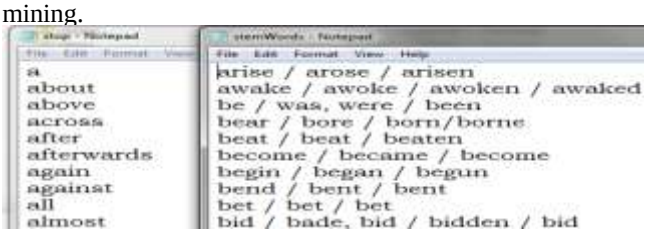

Fig 5: Stop words and Stemmed words

E. *TF Calculation*

Based on the contents, the emails will be classified under spam or ham category. Text mining will be performed on the text of text based mail and also on the extracted text of image based mail. The TF(Term Frequency) of each word in the mail will be calculated and the word with the highest TF will be taken into account. Term Frequency is the frequent measure of the occurrence of a term in a document. It is possible that a term would appear much more times in long documents than shorter ones, since every document is different in length. Thus, the term frequency is often divided by the document length.

F. *Weight Measure*

A weight document is maintained where the weights are assigned in between 0 to1 for the spam words and 0 to -1 for the non-spam words. Then the actual weight of each word will be:

$$\text{Actual weight (t)} = \text{Weight (t)} * \text{TF (t)} \quad (1)$$

G. *Classification using ANN*

The supervised learning is employed to train the neural network. The inputs are provided to the network for which there is a known answer. Hence the network can find out whether it has made a correct guess or not. If the incorrect guess is made, the network can learn from its mistake and adjust its weights.

With the help of sign activation function, the output will be either -1 or 1. The input data will be classified according to the sign of the output. The output function will be:

$$\text{Output} = w1*t1 + w2*t2 + w3*t3 \ldots\ldots\ldots wn*tn + \text{bias} \quad (2)$$

If the output is positive it will be classified as +1 and the negative output will be classified as -1.

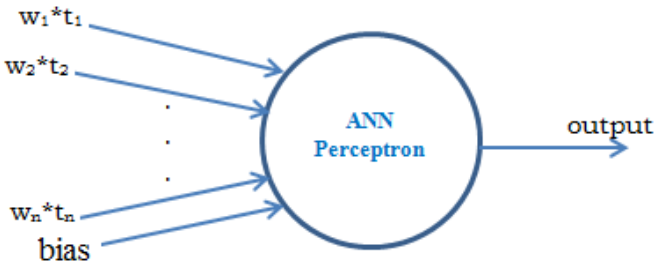

Fig. 6 ANN perceptron

Bias will have the fix value between 0 to 1.

The error can be defined as the difference between the desired answer and its guess.

$$\text{Error} = \text{Desired output} - \text{Guessed output} \quad (3)$$

Here error is the determining factor to adjust the perceptron's weights. If the perceptron's guessed answer equals the desired answer then the error is 0. If the desired answer is -1 and guessed is +1, the error is -2. Similarly if the desired answer is +1 and guessed is -1, then the error is +2.

V. DATASET

In the proposed system mails are taken from the standard data set and also from the personal corpus. SpamAsssassin [12] is the open source dataset which contains 700 spam mails of various categories. SpamArchieve [13] is another open source dataset which contains 1000 image spams of various categories.

VI. CONCLUSION

This paper addresses a problem of the spam emails and offers a solution to detect this problem. In this technique a neural network will be train to be able to distinguish between spams and legitimate emails. The mails could be text based or image based. OCR engine extracts text from image. Once the text is extracted from image the classification process will be same as the text based mail. This approach can efficiently perform classification of spam and ham mails.

REFERENCES

[1] Harisinghney A. ; Dixit A. ; Gupta S. ; Arora A. "Text and Image based spam email classification using KNN, naïve Bayes and Reverse DBSCAN algorithm" Optimization, Reliability and Information Technology(ICROIT) , 2014 International Conference on DOI:10.1109/ICROIT. 2014. 6798302, page(s):153-155, 2014

[2] Sivanadyan, Thiagarajan "Spam? Not Any More! Detecting Spam emails using neural networks" International Conference on Machine Learning and Computing, June-July 1999

[3] James Clark, Irena Koprinska, and Josiah Poon "A Neural Network Based Approach to Automated E-mail Classification " Web Intelligence, IEEE/WIC International Conference , Page(s): 702 – 705, 2003.

[4] N. Nhung and T. Phuong. "An Efficient Method for Filtering Image-Based Spam E-mail". Proc. IEEE International Conference on Research, Innovation and Vision for the Future 10.1I09 /RIVF. 2007.369I41.

[5] Xiao Mang Li, Ung Mo Kim "A Hierarchical Framework for Content-based Image Spam Filtering" International Conference on Machine Learning and Computing, 2010

[6] Sarab M. Hameed1, Noor Alhuda J. Mohammed2 "A Content based Spam Filtering Using Optical Back Propagation Technique" International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 7, July 2013, ISSN 2319 – 4847.

[7] Ms.D.Karthika Renuka1, Dr.T.Hamsapriya2, Mr.M.Raja Chakkaravarthi3 ,Ms. P. Lakshmi Surya4 "Spam Classification based on Supervised Learning using Machine Learning Techniques" , Process Automation, Control and Computing (PACC), Page(s): 1 - 7 IEEE 2011

[8] R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. International Conference on Document Analysis and Recognition, 2007

[9] M. Soranamageswari and Dr. C. Meena "Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks" Second International Conference on Machine Learning and Computing, 2010

[10] Ms. D. Katrina Renuka and Dr.T.Hamsapriya "Spam Classification based on Supervised Learning using Machine Learning Techniques" Process Automation, Control and Computing (PACC), 2011 International Conference on DOI: 10.1109/PACC.2011.5979035, Page(s): 1-7, 2011.

[11] Liu, G., & Yang, F. "The application of data mining in the classification of spam messages" In Computer Science and Information Processing (CSIP), 2012 International Reverse Conference on (pp. 1315-1317), IEEE 2012.

[12] http://spamassassin.apache.org/publiccorpus

[13] www.cs.jhu.edu/~mdredze/datasets/image_spam/spam_archive.tar.gz