Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-5, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Spam Mail Detection Using Artificial Neural Network

[1]Harshal Deshmukh, [2]Chetan Nandeshwar, [3]Sagar Wanjari, [4]Pankaj Bhardwaj, [5]Devendra Ramtekkar & [6]Rajesh Nasare
[1,2,3,4,5,6]Department Of Computer Science and Engineering, RGCER, Nagpur

Abstract: *Today e-mail is the most popular and financially cheapest way of communication for internet users..e-mail is going to be misused due to its popularity. One such misuse is the posting of money offering message, unwanted e-mails known as spam or junk e-mails. E-mail spam has various consequences like productivity is reduced, takes extra memory space in mail boxes, extra time for suffering, software damaging viruses, and materials that have potentially harmful information for Internet users, destroy stability of mail on servers, resulting users to spend lots of time for sorting incoming mail and deleting unwanted correspondence mails. So there is a need of spam detection system so that its consequences can be reduced. In this project our prime aim is to detect text as well as image based spam to achieve the objective we applied ANN algorithm , Pre-processing of email text before executing the algorithms is used to make them predict better .We uses Enron corpus's dataset of spam emails.*

Keywords: *ANN (Artificial Neural Network); stemming; OCR .*

1. Introduction

As number of internet users is increasing day by day, more people are finding email communication an inexpensive way to send their data or information and communicate with their cal legs . With some pros there is also some cons. Almost now every website ask for email id for completing their registration processing, thus making internet users more prone to get affected by the spam mails. This is evident from the fact that spam emails have accounted for 68.8% of all email traffic in year 2012.The increasing numbers of spam emails wastes one's time for deleting such mails and also wastes network resources significantly. Most important they expose users to scams such as phishing and virus attacks .Spammers have now gone a step ahead and to prevent spam filters from detecting their mails, images containing the spam text are sent. This has increased the burden to detect these manifold spam emails.

Spam arises from an online social situation and now a days it become a social problem as well .Spam messages are causing serious problems which overflow our email boxes. Most email readers spend a non-trivial amount of time regularly by deleting such junk email messages, even as an expanding volume of such email occupies server storage space and consumes network bandwidth. Due to this problem it becomes necessary to distinguish between legitimate and spam emails. Although using some anti-spam technologies are applied for successful filtering text based spam emails. The image spam is substantially more difficult to detect, as they contain a variety of image creation and image randomization algorithms .In the Image spam contain the text message which is embedded into attached images to defeat the anti-spam filters technique . The basic rationale behind image spam is that it is difficult to find such image spam using spam filtering software designed to detect patterns in image text in the plain-text email body.

- **Experimental Work**
- **Extract words from Image**

To extract the text from image is an arduous task. It must be done by sophisticated OCR tool sand based on the high level, low level, and combination of both the features of image in a spam mail can be predicted. All those web pages and domains that are notorious for sending spam mails and are not trusted; go on the list of black list .Thus, if a domain that matches from this list, the mail is predicted spam without any further processing. Further, spam is in the eye of the recipient, so a white list is maintained where users can mark those websites they want mails from whether they send "spam" or not. Thus no processing is done when a white listed domain matches.

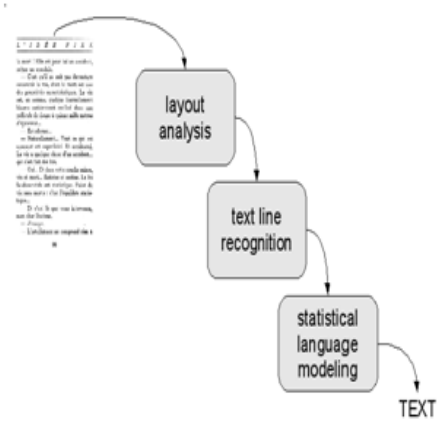

Fig1. Flow diagram of OCR engine

2. Dataset

A large set of email messages, form the Enron corpus, which was made public by the legal investigation concerning the Enron Corporation. Enron corpus contains a total of messages belonging to users with an average of messages per user. This is approximately one third the size of the original corpus.

3. Methodology

Classification

The learning of artificial neural network is fully unsupervised. And the Name suggest it is simply a special case in which there are only two classes. Binary classification is the process which classify the given elements of set into two group on the basis of a classification rule. Now there are various binary classifiers for learning paradigms are 1. Decision Tree
2. Neural Networks
3. Bayesian classification
4. SVM.

Neural Network

A neural network is a set of connected input or output units connection has a weight associated with it Back propagation is a neural network learning algorithm. Layer is made up of units. For each training tuple, the inputs to the network correspond to the attributes measured. The inputs are fed simultaneously in the units making up the input layers .After which these inputs are passed through the input layer which then weighted and fed simultaneously to a second layer of hidden layer "neuron like" units. The hidden layer output units can be input to another hidden layer, and so on.

Unsupervised Learning

This is required when there is not an example of data set with known answers. Imagine we are searching a hidden pattern contain in a data set

Proposed System

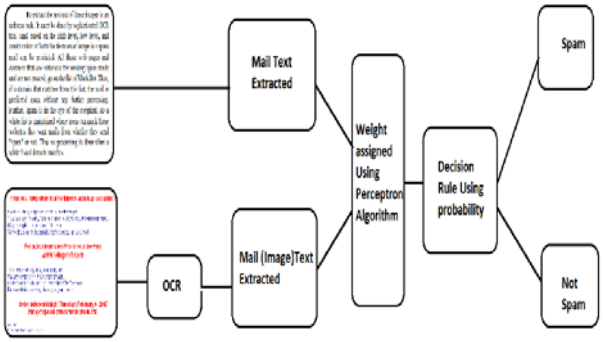

Fig3. Proposed System

The proposed System mainly concentrate on the body of email which generally contain spam words. In this words two dataset are considered. In which one is Spam dataset of 500 sorted mails having Mixture of Spam and non Spam mails another is a pure spam Archives dataset having 700 Spam Images. The Email Contained are analyzed and the list of words are weighed according to the probability of being a spam oriented words.

Black List and White List

Black Listing: Black-listing is creating a list of domain names which are used by the spammers, when a mail comes from spammer specific domain which is in black list, it is directly considered as spam. No further processing is done.

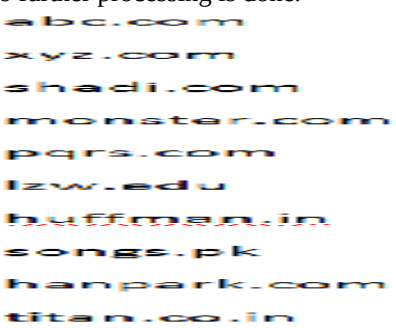

Fig4. Black List

White Listing: White list is a list of trusted domains. White listing method used to classify user's email addresses as legitimate ones. But blacklisting and

white listing is not always accurate. Therefore, to counter employed by spam filters apply all these techniques, spammers used to send mails with embedded images containing the spam text. Extracting a text from these images is an arduous task. It must be done by sophisticated OCR tools and based on the high level, low level, and combination of both the features of image in a spam mail can be predicted.

4. Probability calculation:

Frequently occurring term in a document is calculated by this. It may possible that a term in the document appear more time for a long or shorter document, Every document have different length.

```
Probability of Offer =0.025768645864

Probability of sales =0.005285654656

Probability of Rs = 0.596618426456

Probability of Health = 0.56876624162

Probability of family = 0.56643987898

Probability of money = 0.69999478555

Probability of save = 0.22235556888

Probability of Free = 0993565555
```

Fig5. Calculated Probability

Weight calculation

A weight document is important part of our system. The fully updated weight document can make system stronger. The weight are assigned in between 0 to 1 for spam words and for non spam word weight is assign between 0 to -1.

```
sale=0.75
save=0.3
offer=0.60
free=0.5
adclick =1.0
http=0.1
insurance=0.2
country=-0.5
youll=-0.1
disregard=-0.4
www=0.1
com=0.1
net=0.1
org=0.1
in=0.1
co=0.1
pk=0.1
uk=0.1
jp=0.1
ie=0.1
period=0.1
```

Fig6. Weight assigning

Thus, the actual weight will be calculated as :

ACTUAL_WEIGHT(T)=WEIGHT(T)* PROBABILITY(T)

Where T is the term considered.

5. Conclusion

In this project we applied ANN algorithms to detect spam mails.we are using 250 spam mails dataset for training and testing the algorithms. GUI Design using Net Beans have been completed.

Various spam detection steps such as pre-processing step or data cleaning step (like stop words removal, stemming), representation of data, and classification of spam or non-spam e-mail messages are discussed.

6. References

[1]Nosseir,Khaled Nagati and Islam Taj-Eddin Intelligent, ”Word based totally Spam Filter Detection exploitation Multi-Neural Networks”, IJCSI International Journal of engineering issues Egypt, 1,March,2013,Vol. 10, Issue 2.

[2] Anirudh Harisinghaney, Arnan Dixit, Saurabh Gupta, Anuja Arora, “Text and Image based totally Spam Email Classification exploitation KNN, Naive Bayes and Reverse DBSCAN Algorithm”, International Conference on reliableness, improvement and information Technology,Date: Gregorian calendar month 6-8 2014.

[3] C. Gulyás, "Creation of a theorem network-based totally meta spam filter, exploitation the analysis of varied spam filters", Master Thesis, Budapest, sixteenth would possibly 2006.

[4]Enron corpus data http://archive.ics.uci.edulmIldatasets/Spambase, http://spamassassin.apache.org/publiccorpus/

[5] Spamhaus, “Definition of Spam”,Date:8, Feb,2010. http://www.spamhaus.org/definition.html

[6] Ling-Spam info set. || Internet: http://csmining.org/index.php/ling-spam-datasets.html.

[7] Firte, L., Lemnaru, C. and Potolea, R. 2010,”Spam Detection Filter exploitation KNN rule and Resampling”, Intelligent laptop computer Communication and method (ICCP), 2010 IEEE International Conference , pp. 27 – 33.

[8] Mr. Rahul Bansod, Mr. R. S. Mangrulkar Ms. V. G. Bhujade,”Spam Classification exploitation ANN with

Weight live ", International Journal of Advance laptop computer Technology(IJACT), ISSN: 2319-7900

[9] W.A. Awad1 and S.M. ELseuofi,"Machine Learning methodology for Spam E-Mail classification",International journal of engineering And information technology(IJCSIT),Date. Feb-2011, vol-3, no.1

[10] Ismaila Idris ,"E-mail spam classification with Artificial Neural Network and Negative alternative Algorithm", International journal of engineering And Communication Network (IJSCSCN) ,vol-1(3), 227-231 ISSN: 2249-5789