International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

Spam Classification using Artificial Neural Network with Weight Measures

Mr. Rahul Bansod,
BDCOE, Sewagram
rahulbansod15@gmail.com

Mr. R. S. Mangrulkar
BDCOE, Sewagram
rsmangrulkar@gmail.com

Ms. V. G. Bhujade
BDCOE, Sewagram
vaishali.hardeo@gmail.com

Abstract

Email is one of the instant forms of business communication. As the usages of email continue, unrequested bulk emails called spam emails also growing continuously. Millions of spam mails are sent over internet every day to targeted population to advertise services and products. These annoying unsolicited Commercial emails occupy storage space in the server and consume network bandwidth in large amount. Recently image spam becomes widespread and does a lot of harm. It is observed that one out of every six emails is a spam mail and this rate is growing rapidly. The effective removal of these spam mails is mandatory. Spam emails need to be classified and separated from ham mails, as they are the source of financial loss and annoyance for the recipients. Most of the supervised machine learning methods uses either header-based or content-based features for the classification of spam mails. In this paper we present a technique to classify both text and image spam mails using Artificial Neural Network with negative and positive weight measures. OCR tool is used to extract text data from the image.

Keywords: ANN (Artificial Neural Network), Image spam, OCR.

Introduction

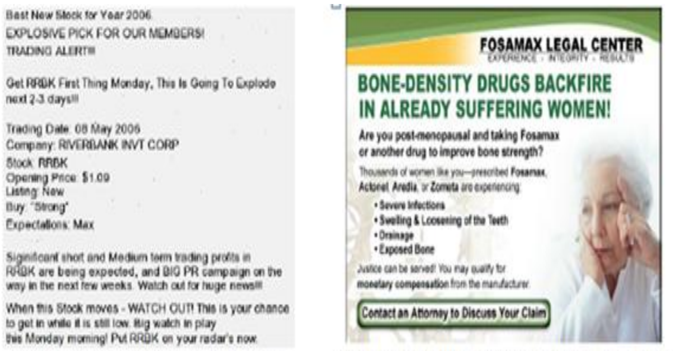
In the recent years the mass of undesired mails called spam mails has increased fabulously. These often considered junk. mails or junk postings. Spams are mostly the advertisements for some product sent to a mailing list or newsgroup .

In addition spam also consumes a lot of network bandwidth. Consequently, there are many organizations, who have taken it upon themselves to resist strongly against spam with a variety of techniques. But there is truly negligible that can be done to prevent spam, because the Internet is public. However, some policies are undertaken to prevent spammers from spamming their infuriating advertisements.

To detect spam based on the textual content of the email, many text-based anti-spam approaches have been proposed, such as Bayesian filters and Support Vector Machine (SVM) filters. However, these approaches soon lost their effectiveness because spammers have introduced a trick to embed junk information into images. See the example of normal image spam shown in Fig.1. Anti-spam filters are posing a great challenge in detecting these spam emails where advertisement text embedded in images [1]**.** Text-based spam filtering techniques are failed to detect the spammer's new approach. Spammer's advertisements have become a part of an embedded image file attachment rather than the body of the e-mails.

Statistics says, up to 25% of spam being sent today contains images and this number is gradually increasing. Therefore, to detect image-based spam it is desirable to develop systems. One of the possible ways to detect image-based spam is a pipeline of an optical character recognition (OCR) system [11] that extracts text embedded in the image, followed by a text classifier to separate advertising text from genuine content.

Accuracy, flexibility and speed are the main features that characterize a good OCR system. Several algorithms for character recognition have been developed based on feature selection. The performance of the systems has been constrained by the dependence on font, size and orientation. The recognition rate in these algorithms depends on the choice of features. Most of the existing algorithms involve extensive processing on the image before the features are extracted that results in increased computational time.



Best New Stock for Year 2006
EXPLOSIVE PICK FOR OUR MEMBERS!
TRADING ALERT!!!

Get RRBK First Thing Monday, This Is Going To Explode next 2-3 days!!!

Trading Date: 08 May 2006
Company: RIVERBANK INVT CORP
Stock: RRBK
Opening Price: $1.09
Listing: New
Buy: "Strong"
Expectations: Max

Significant short and Medium term trading profits in RRBK are being expected, and BIG PR campaign on the way in the next few weeks. Watch out for huge news!!!

When this Stock moves - WATCH OUT! This is your chance to get in while it is still low. Big watch in play this Monday morning! Put RRBK on your radar's now.

(a) Image containing only text

FOSAMAX LEGAL CENTER
EXPERIENCE · INTEGRITY · RESULTS

BONE-DENSITY DRUGS BACKFIRE IN ALREADY SUFFERING WOMEN!

Are you post-menopausal and taking Fosamax or another drug to improve bone strength?

Thousands of women like you—prescribed Fosamax, Actonel, Aredia, or Zometa are experiencing:

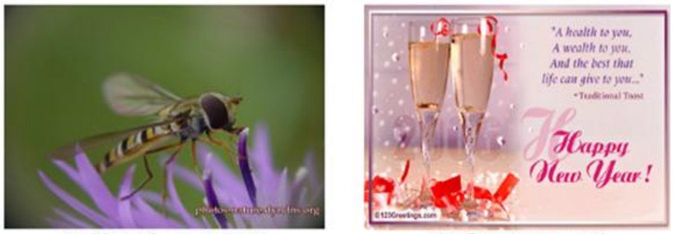
- Severe Infections
- Swelling & Loosening of the Teeth
- Drainage
- Exposed Bone

Justice can be served! You may qualify for monetary compensation from the manufacturer.

Contact an Attorney to Discuss Your Claim

(b) Image with photographic elements

Figure 1. **Spam Images**



(a) Natural scene photo

(b) Greetings e-card

Figure 2. **Normal images**

Related Work

Today's internet world is facing many challenges for image spam classification [5]. Many algorithms have been proposed for classifying spam and legitimate emails [6]. The fidelity of the spam classifiers introduced by service providers are handled by the spammers through various randomization techniques. However many spam classification approaches have been put forth by the researchers to assist the developers of anti-spam detectors.

Liu Yang et al. [7] proposed a scheme for filtering the image spam based on the method of spam behavior recognition filtering. A model is developed based on Bayes technique which identifies spam according to the behavior of mail sent. This approach adopts filtering of spam by several stages, by utilizing the least risk of Bayes technique in order to recognize the image spam as early as possible.

Ngo Phuong et al. [8] proposed a method which uses an edge-based feature vector. These vectors can be computed efficiently, to represent major shape properties of the image instead of extracting embedded text from an image. A vector of similarity measures is computed from an image to a small set of gold standards. These similarity vectors are then served as input for SVMs training and classification. Due to less expensive image processing and text recognition steps this method is fast.

Chao Wang et al. [9] proposed a two stage method for a feature extraction scheme which concentrates on low-level image features such as Type, Size, Width, Height and Bit Depth of image, which can perform classification rapidly. First by getting image features, which contains file properties, color features and texture properties, it runs a one-class SVM classifier with RBF kernel to classify image spam in the second stage.

Xiao Mang Li et al. [10] proposed a hierarchical anti-spam framework, which adopts multiple techniques. It includes text classification, image processing and Optical Character Recognition in different layers which consist of filtering module and data module. The four filtering layers are formed in a hierarchical manner and are managed by the controlling unit. Each layer is capable of making filtering decisions. The data module provides data support for decision making process. Text localization and obscuring detection are utilized to improve the recognition rate of the OCR tool and to detect obscured image spam.

The Pitfalls of Spam Emails

Email is viewed as the widest and most convenient application for transferring message on Internet. With the fast development of Email services, spam messages are increasing rapidly, and the contents of spam messages are related to all aspects of life. The numbers are very large. The problems of spam messages has been seriously disrupting people's normal work and life. Spam messages brought a very bad influence on social harmony.

- Biggest nuisance on the net which affect the social stability

Spam messages are causing serious problems which flood our email boxes in quick time. Companies are wasting much time on detecting and removing spams.

- To interfere with normal communication

Spam messages can be send in mass, so the transmission time will take up network bandwidth, causing congestion, affecting the performance of the network and people's normal communication.

- Targeting Social sites

Blogging software (Blog) places comments repeatedly to various blog posts openly. Spammers take advantage of this open nature of comments and provide links to the spammer's commercial web site. The most common technique involves posting pornographic links. Online dating can be fixed on the comments section of profiles. Bots to post messages is another repeatedly used technique along with enticing text and images on random user's profiles, consistently sexually suggestive messages.

Dataset

Two datasets are used in the system. SpamAsssassin [11] and SpamArchieve [12] . For text based spam detection, SpamAsssassin is the open source dataset which contains 700 text based spam mails of various Category. SpamAssassin has a modular architecture that allows other technologies to be quickly wielded against spam and is designed for easy integration into virtually any email system. It is an intelligent email filter which uses a diverse range of tests to identify unsolicited bulk email.

For image based spam detection, SpamArchieve [10] is the open source dataset. The SpamArchive dataset was randomly downloaded from the SpamArchive website. In the SpamArchive dataset, 25000 spam images. Many emails did not have images explicitly attached, but they provided links to the images.

Methodology

The methodology of spam classification is based on the figure below:

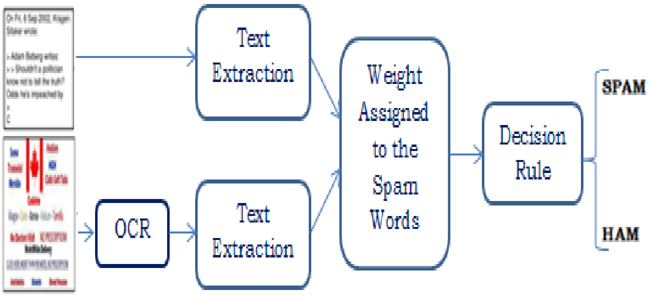
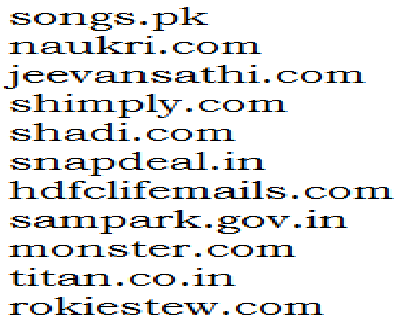


Figure 3. Proposed System

The proposed method mainly concentrates on the body of the email which generally contains spam words to target the users. In this work two datasets are considered. One is a spam assassin dataset of 500 mails having mixture of spam and non-spam mails and another is a spam archives dataset having 700 spam images. The email contents are analyzed and the lists of words are weighted according to their probability of being a spam oriented word. Based on the value of cognitive load we differentiate between spam and ham mails.

A. Black Listing and White listing

All those web pages and domain that are widely known for sending spam mails and are not trusted, go onto the black list. If a domain that matches from this list, the mail is predicted spam without any further processing.

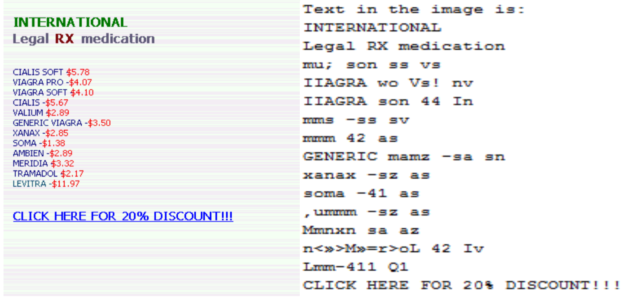


songs.pk
naukri.com
jeevansathi.com
shimply.com
shadi.com
snapdeal.in
hdfclifemails.com
sampark.gov.in
monster.com
titan.co.in
rokiestew.com

Figure 4. Black List

B. Words Extraction from Image

Image spam is a kind of email spam where the message text of the spam is presented as a picture in an image file. The image is passed through the google's open source library Tesseract, and words are extracted from it. Optical Character Recognition, or OCR, is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data. In OCR processing, the bitmap is analyzed for light and dark areas in order to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code.



INTERNATIONAL
Legal RX medication

CIALIS SOFT $5.78
VIAGRA PRO -$4.07
VIAGRA SOFT $4.10
CIALIS -$5.67
VALIUM $2.89
GENERIC VIAGRA -$3.50
XANAX -$2.85
SOMA -$1.38
AMBIEN -$2.89
MERIDIA $3.32
TRAMADOL $2.17
LEVITRA -$11.97

CLICK HERE FOR 20% DISCOUNT!!!

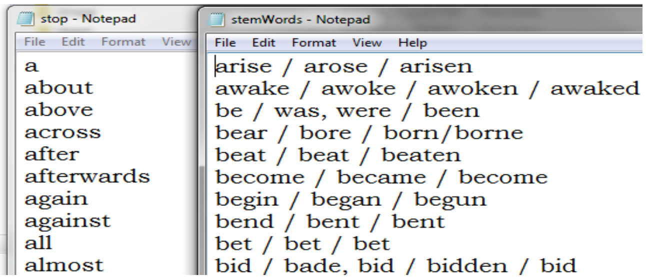
```
Text in the image is:
INTERNATIONAL
Legal RX medication
mu; son ss vs
IIAGRA wo Vs! nv
IIAGRA son 44 In
mms -ss sv
mmm 42 as
GENERIC mamz -sa sn
xanax -sz as
soma -41 as
,ummm -sz as
Mmnxn sa az
n<»>M»=r>oL 42 Iv
Lmm-411 Q1
CLICK HERE FOR 20% DISCOUNT!!!
```

Figure 5. Image spam & extracted text from image

C. Preprocessing of Data

Stemming is the process of moving any word to its root value. 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.



a
about
above
across
after
afterwards
again
against
all
almost

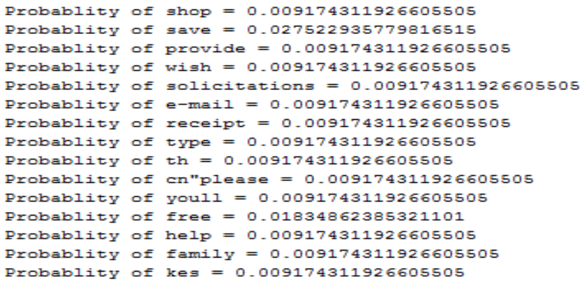
arise / arose / arisen
awake / awoke / awoken / awaked
be / was, were / been
bear / bore / born/borne
beat / beat / beaten
become / became / become
begin / began / begun
bend / bent / bent
bet / bet / bet
bid / bade, bid / bidden / bid

Figure 6. Stop words and Stemmed words

Many words in natural language occur with high frequency but have low information content, such as ”a,” ”an,” ”the,” and most prepositions and conjunctions can be removed with the assumption that no serious loss of information will occur. These so called stop words are specified in a list and can be removed from the token stream.

D. Probability Calculation

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.

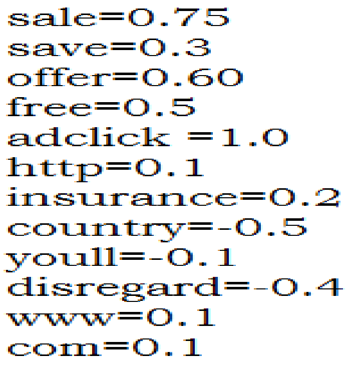


```
Probablity of shop = 0.009174311926605505
Probablity of save = 0.027522935779816515
Probablity of provide = 0.009174311926605505
Probablity of wish = 0.009174311926605505
Probablity of solicitations = 0.009174311926605505
Probablity of e-mail = 0.009174311926605505
Probablity of receipt = 0.009174311926605505
Probablity of type = 0.009174311926605505
Probablity of th = 0.009174311926605505
Probablity of cn"please = 0.009174311926605505
Probablity of youll = 0.009174311926605505
Probablity of free = 0.01834862385321101
Probablity of help = 0.009174311926605505
Probablity of family = 0.009174311926605505
Probablity of kes = 0.009174311926605505
```

Figure 7. Calculated Probability

E. Weight Measure

A weight document is the heart of our system. The fully updated weight document can make this system stronger. The weights are assigned in between 0 to1 for the spam words and 0 to -1 for the non-spam words.



```
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
```

Figure 8. Assigned Weights

Thus the actual weight of each term will be calculated as:

ACTUAL_WEIGHT (T) = WEIGHT (T) * PROBABILITY (T) (1)

Where T is the term considered

F. Classification using ANN

The learning of Artificial Neural Network fully supervised. The inputs are provided to the ANN for which there is a prominent answer. Hence it is required to check whether the network has made a correct guess or not. If the incorrect guess is made, the network learns 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:

SUM= W1*T1 +W2*T2 +W3*T3 ………Wn*Tn + Bias_value (2)

If the sum is positive it is classified as +1 and the negative sum is classified as -1.

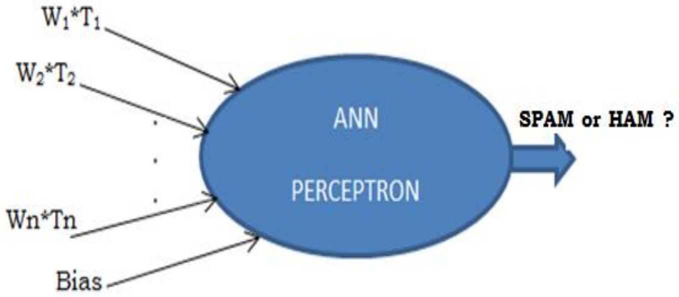
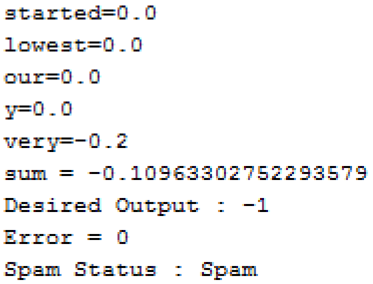


Figure 9: ANN perceptron

The Bias value is taken between 0 to 1. It is fixed to 0.5.

Error = Desired_value – Guessed_value (3)

Here to adjust the perceptron's weight error is the determining factor. The error is 0 if the perceptron's guessed answer equals the desired answer. The error is -2 if the desired answer is -1 and guessed is +1. Similarly the error becomes +2 if the desired and guessed answer is +1 and -1 respectively.



```
started=0.0
lowest=0.0
our=0.0
y=0.0
very=-0.2
sum = -0.10963302752293579
Desired Output : -1
Error = 0
Spam Status : Spam
```

Figure 10. Sample Output

Conclusion

In this paper, we have presented a method to filter text spam and image spam by assigning negative and positive weights to the words based on their probability of being promotional or non-promotional word. To fight against spam is a difficult problem for the entire world. The advent of image spam is one more challenge for the majority of internet users. Image spam erodes the limited network resources, and brings troubles to people. The war between spam and anti-spam filters is growing rapidly. Therefore, we should search for the preventive measures, and at the same time we should also predict the evolution trend of spam as the image spam is not the last variant of spam.

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] J. D. Brutlag and C.Meek, “Challenges of the email domain for text classification,” in ICML, 2000, pp. 103–110.

[3] 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 (RIVF07), IEEE Press, Mar. 2007 , pp. 96-102. doi: 10.1I 0 9 /RIVF.2007.369I 4 1.

[4] R. Smith, "An Overviewof the Tesseract OCR Engine", in Proc. International Conference on Docment Analysis and Recognition, 2007.

[5] 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 .

[6] 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, 2011, Page(s): 1 – 7.

[7] 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.

[8] 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.

[9] Chao Wang, Fengli Zhang, Fagen Li, Qiao Liu ”Image Spam Classification based on low-level Image Features” Communications, Circuits and Systems(ICCCAS), 2010, Page(s):290-293, IEEE.

[10] Xiao Mang Li, Ung Mo Kim “A Hierarchical Framework for Content-based Image Spam Filtering” Information Science and Digital Content Technology (ICIDT), 2012, 8th International Conference, Page(s): 149-155, IEEE.

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

[12] www.cs.jhu.edu/~mdredze/datasets/ image_spam/ spam_ archive.tar.gz