INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue3)
Available online at www.ijariit.com

E-Mail Spam Detection and Classification Using SVM and Feature Extraction

Shradhanjali
Rungta College of Engineering and Technology
Dept. of Computer Science and Engineering
Bhilai, Chhattisgarh, India
shradhanjali.nirmal24@gmail.com

Prof. Toran Verma
Rungta College of Engineering and Technology
Dept. of Computer Science and Engineering
Bhilai, Chhattisgarh, India
toran.verma@rungta.ac.in

Abstract: Today emails have become to be a standout amongst the most well-known and efficient types of correspondence for Internet clients. Hence because of its fame, the email will be misused. One such misuse is the posting of unwelcome, undesirable messages known as spam or junk messages. Email spam has different consequences. It diminishes productivity, consumes additional space in mailboxes, additional time, expands programming damaging viruses, and materials that contain conceivably destructive data for Internet clients, destroys the stability of mail servers, and subsequently, clients invest lots of time for sorting approaching mail and erasing undesirable correspondence. So there is a need for spam detection so that its outcomes can be reduced. In this paper, propose a novel method for email spam detection using SVM and feature extraction which achieves an accuracy of 98% with the test datasets.

Keywords: Spam, Types of Spam, Email Spam, Classification, SVM.

I. INTRODUCTION

spam refers to unsolicited business email. Otherwise called junk mail, spam floods Internet client's electronic mailboxes. These junk emails can contain different sorts of messages, for example, commercial advertising, pornography, business promoting, doubtful product, infections or quasi-legal services [3].

A. Types of Spam

Fundamentally, spam can be classified into the accompanying four types:

- Usenet Spam
- Texting Spam
- Mobile Spam
- E-mail Spam

Usenet Spam: User Network is an open get to arrange on the Internet that gives group talks and group email informing. All the data that goes over the Web is called "NetNews" and a running accumulation of messages about a specific topic is known as a "newsgroup". Usenet spam is presenting of some commercial on the newsgroups. Spammers focus on the clients those read news from these newsgroups. Spammers present promotion on a substantial measure of newsgroups at once. Usenet spam rob clients of the utility of the newsgroups by overwhelming them with a barrage of promoting or other unrelated posts.

Instant Messaging Spam: Instant informing frameworks, for example, Yahoo Messenger, AOL Instant Messenger (AIM), Windows Live Messenger, Facebook Messenger, XMPP, Tencent QQ, Instant Messaging Client (ICQ), and MySpace talk rooms are all objectives for spammers. A few IM frameworks give a registry of clients, including statistic data, for example, date of birth and gender. Advertisers can gather this data, sign on to the framework, and send undesirable messages, which could incorporate business malware, viruses, and associates to paid destinations [8]. As texting has a tendency to not be stuck by firewalls;

subsequently, it is a particularly helpful route for spammers. It focuses on the clients when they join any visiting space to discover new friends. It ruins appreciate of individuals and wastes their time moreover.

Mobile Phone Spam: Mobile phone spam is focused on the content informing administration of a cell phone. This can be particularly irritating to clients not just for the bother additionally in light of the cost they might be charged per instant message gotten in a few markets. This sort of spam more often than not contains a few plans and offers on different items. In some cases, service providers likewise make utilization of this to trap the client for activation of some paid services.

Email Spam: Email spam is the most well-known type of spam. Email spam focuses on the individual clients with direct emails. Spammers make a rundown of email clients by inspecting Usenet postings, stealing lists of webmail, search the web for e-mail addresses. Email spam costs cash to a client of email in light of the fact that while the client is perusing the messages meter is running. Email spam additionally costs the ISPs on the grounds that when a majority of spam sends are sent to the email clients its waste the bandwidth of the service providers these expenses are transmitted to clients. All undesirable emails are not spammed messages.

B. Classification

An extensive number of classification algorithm has been connected to spam recognition region, where support vector machine classification for its decent generalization performance effect Furthermore, exceptionally well known. SVM is an intense method utilized for data classification. Despite the fact that people consider that it is simpler to use than Neural Networks. Each example in the preparation set contains one class marks and a few components. The fundamental point of SVM is to create a model which predicts class labels of information occurrences in the testing set which are given only the features. At the show, the support vector machines have been broadly utilized as a part of content based hostile to spam system. SVM is a splendid solution for the little sample size issue, by developing an isolating hyperplane to finish the classification. As the support vector machine in spam identification in the great execution, the paper utilizes this algorithm to identify spam emails.

1) Support Vector Machine

A support vector machine (SVM) can be utilized when our information has totally two classes. An SVM classifies information by finding the ideal hyperplane that isolates all information purposes of one class from those of alternate class. The hyperplane for an SVM implies the one with the biggest margin between the two classes. Margin implies the maximal width of the segment parallel to the hyperplane that has no interior information points.

2) Properties of SVM

Support Vector Machine has a place with a group of generalized linear classifiers and it can be deciphered as an extension of the perception. A unique property is that they all the while limit the exact classification error and maximize the geometric margin; henceforth they are otherwise called maximum margin classifiers.

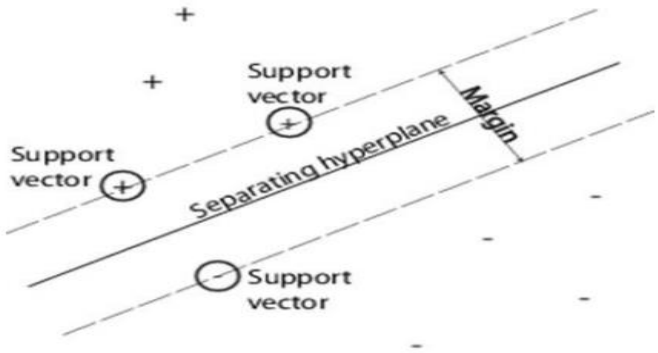

Fig 2: Support Vector Machine

II. LITERATURE SURVEY

MD. Rafiqul Islam et al. [13] discussed different machine learning algorithms for spam filtering and presented a comparative study of spam filters. Their research includes a study of automated filtering and machine learning techniques like rule based, content based, personalized, collaborative, support vector machine and kernel-based algorithms for filtering spam.

Ni Zhang et al. [14] developed a method for filtering spam emails from the Internet service providers in its heavy traffic. They applied their method to email traffic data captured at one of the largest commercial Internet service providers in China. They achieved a result of 70.4% reduction of junk mail traffic.

Seongwook Youn et al. [15] proposed a comparative study for email classification. Neural Network, SVM, Naive Bayesian and J48 classifiers are used to filter spam from the datasets of emails. A neural network consists of data preprocessing, data training and testing.

Enrico Blanzieri et al. [16] proposed a survey on learning based techniques of spam filtering. This Paper discussed the learning based methods of spam filtering like keyword filtering, image based filtering, and language based filtering, filters based on non-content features, collaborative filtering and hybrid approaches.

A.G.Lopez-Herrera et al. [17] developed a multi-objective evolutionary algorithm for filtering spam. They evaluated the concepts of dominance and Pareto set. SPAM-NSGA-II-GP is used for filtering spam emails. MOEA is used to learn a set of queries with good precision and recall. PUI datasets are used for spam filtering.

METHODOLOGY

In this section, we will discuss the proposed methodology for email spam detection technique. The fig. 1. Shows the workflow.

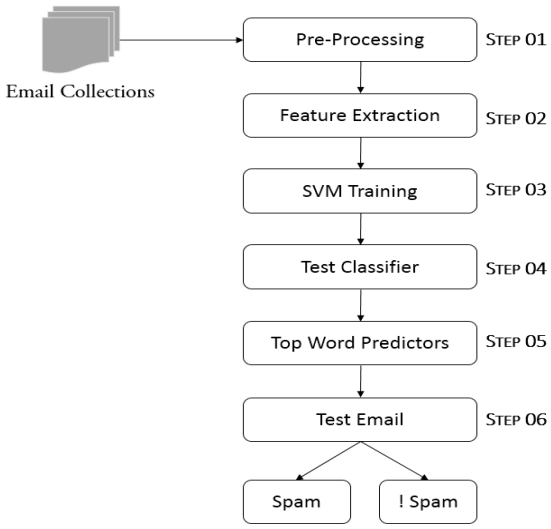

Fig. 1: Proposed Workflow Architecture

A. Preprocessing

The pre-processing step is used to remove the noises from the email which are irrelevant and need not be present. The preprocessing step includes.

- Removal of Numbers
- Removal of Special Symbol
- Removal of URLS
- Stripping HTML
- Word Stemming

B. Feature Extraction

Feature Extraction is used to extract the important and relevant features from the email body. The feature transforms the email into 2 D vector space having features number. These features are mapped from the vocabulary list.

$$x = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 1 \\ 1 \\ . \\ . \\ 0 \end{bmatrix} \in \mathbb{R}^n$$

C. SVM Training

The email spams are used for the training purposes. The training dataset contains spam content and classifier are trained using it. After training, the classifier is ready to classify the spam emails.

D. Test Classifier

The classifier is tested with numerous training data to test the accuracy of the classifier. The proposed solution achieves up to 98 % accuracy in classifying emails.

E. Test Email

After the training phase is completed, a sample email is given as input to the classifier to classify the email. The classifier produces output in the forms of 0 or 1, 1 means it is spam and 0 means it is not a spam.

III. RESULTS AND DISCUSSION

In his section detail result is explained with each stage output. MATLAB is used for executing the algorithm. Below figures present output of each phase.

A. Dataset Description

The email dataset is used for classification of spam and non-spam emails. The dataset is taken from Apache public corpus [13]. The dataset is in the format as shown below.

```
> Anyone knows how much it costs to host a web portal ?>Well, it
depends on how many visitors you're expecting.This can be
anywhere from less than 10 bucks a month to a couple of $100.
You should checkout http://www.rackspace.com/ or perhaps
Amazon EC2 if youre running something big..To unsubscribe
yourself from this mailing list, send an email to:groupname-
unsubscribe@egroups.com
```

Fig.2. Dataset Snapshot

B. Framework Output

Preprocessing:

In the pre-processing step all the number, special symbol are removed and in place of that, some text is placed. The URL and HTML tags are also removed. Word stemming is done to remove unnecessary alphabet from the words.

```
==== Processed Email ====

anyon know how much it cost to host a web portal well it depend on how mani
visitor you re expect thi can be anywher from less than number buck a month
to a coupl of dollarnumb you should checkout httpaddr or perhap amazon ecnumb
if your run someth big to unsubscrib yourself from thi mail list send an
email to emailaddr
```

Fig. 3. Pre-processing Step

Feature Extraction

After pre-processing, the feature is extracted. The main concept behind extraction of feature that the matched word from the dictionary is extracted and are mapped. The mapping is done via using Vocab file.

```
=========================
Length of feature vector: 1899
Number of non-zero entries: 45
```

Fig. 4. Feature Extraction Step

Training

After extracting features, the training is done. In training, the emails are provided as input to the SVM classifier. The SVM classifier takes and generates indices of each word.

```
Training Linear SVM (Spam Classification)
(this may take 1 to 2 minutes) ...

Training .....................................
```

Fig. 5. Training Step

Training and Test Case Accuracy

After training the test case emails are given input to test the accuracy of the system. The accuracy is achieved upto 98% while classifying all the test email.

```
Training Accuracy: 99.850000

Evaluating the trained Linear SVM on a test set ...
Test Accuracy: 98.900000
```

Fig. 6. Training and Test Case Accuracy

Email Spam or !Spam

Finally, the classifier is tested with the given sample email. The class

```
==== Processed Email ====

if you ar a motiv and qualifi individu i will person demonstr to you a system
that will make you dollarnumb number per week or more thi is not mlm

========================

Processed Spam1.txt

Spam Classification: 1
(1 indicates spam, 0 indicates not spam)
```

Fig.7. Email Classification

At the last stage given a sample email, and it is classified as spam or !spam based on its content.

DISCUSSION

In this paper SVM in a combination of feature extraction is used for classification of email spams. The spam can be very harmful and can leave its effect for the long duration of time. It may collapse the whole system. The proposed model can effectively analyzes the email spam and classify them as spam and non-spam. The proposed classifier archives accuracy of 98%.

CONCLUSION

The Spam is a standout amongst the most irritating and malicious increments to worldwide PC world. In this paper, we propose a novel method for email spam detection which can effectively identify the spam emails from its contents. The spam emails can be blocked by the user and genuine mail can be retained by the user. The proposed classifier achieves 98 % accuracy whiles classifying the series of datasets.

REFERENCES

[1] S. Abduelbaset M. However, Tarik Rashed, Ali S. Elbekaie, and Husien A. Alhammi, “An Anti-Spam System Using Artificial Neural Networks And Genetic Algorithms” (A Neural Model In Anti Spam).

[2] Er. Seema Rani, Er. Sugandha Sharma, “Survey on E-mail Spam Detection Using NLP”, International Journal of Advanced Research in Computer Science and Software Engineering, India, Volume 4, Issue 5, May 2014.

[3] Masurah Mohamad, Khairulliza Ahmad Salleh, “Independent Feature Selection as Spam-Filtering Technique: An Evaluation of Neural Network”, Malaysia.

[4] El-Sayed M. El-Alfy, “Learning Methods For Spam Filtering”, College of Computer Sciences and Engineering King Fahd University of Petroleum and Minerals, Saudi Arabia.

[5] Upasna Attri & Harpreet Kaur, “Comparative Study of Gaussian and Nearest Mean Classifiers for Filtering Spam E-mails”, Global Journal of Computer Science and Technology Network, Web & Security, USA, Volume 12 Issue 11 Version June 2012.

[6] Alia Taha Sabri, Adel Hamdan Mohammads, Bassam Al-Shargabi, Maher Abu Hamdeh, “Developing New Continuous Learning Approach for Spam Detection using Artificial Neural Network (CLA_ANN)”, European Journal of Scientific Research, ISSN 1450-216X Vol.42 No.3 (2010), pp.511-521.

[7] Enrique Puertas Sanz, José María Gómez Hidalgo,José Carlos Cortizo Pérez, “Email Spam Filtering”, Universidad Europea de Madrid Villaviciosa de Odón, 28670 Madrid, SPAIN.

[8] Ravinder Kamboj, “A rule based approach for spam detection” ,Computer Science and Engineering Department, Thapar University, India, July 2010.

[9] Vandana Jaswal, Nidhi Sood, “Spam Detection System Using Hidden Markov Model”, International Journal of Advanced Research in Computer Science and Software Engineering, India, Volume 3, Issue 7, July 2013.

[10] Sahil Puri, Dishant Gosain, Mehak Ahuja, Ishita Kathuria, Nishtha Jatana, “Comparison And Analysis Of Spam Detection Algorithms”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), India, Volume 2, Issue 4, April 2013.

[11] Ann Nosseir , Khaled Nagati and Islam Taj-Eddin, “Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks” , IJCSI International Journal of Computer Science Issues, Egypt, Vol. 10, Issue 2, No 1, March 2013.

[12] Jitendra Nath Shrivastava, Maringanti Hima Bindu, “ E-mail Spam Filtering Using Adaptive Genetic Algorithm”, I.J. Intelligent Systems and Applications,MECS, India, January 2014.

[13] http://spamassassin.apache.org/publiccor

