International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

Email Spam Classification Using Support Vector Machine Algorithm

Dhanendra Kumar Dewangan[1], Poonam Gupta[2]
[1, 2]*Central College of Engineering and Management, Dept. of Computer Science and Engineering, Raipur, Chhattisgarh, India*

Abstract: In the present life, internet is a vital part. We invest the vast majority of our time on internet. One of the vital features of internet is communication. Email is a method of correspondence which is utilized for the individual as well as business reason. Spam emails are the emails beneficiary does not wish to take conveyance of; it is moreover called unwanted bulk email. Emails are utilized every day by number of user to speak worldwide. At present vast volumes of spam emails are reasoning genuine inconvenience for Internet user and Internet service. For example, it degrade user examine knowledge, it helps transmission of virus in network, it expands stack on arrange movement. It additionally misuses user time, and vitality for legitimate emails among the spam. Hence, there is a requirement of spam detection so that its result can be reduced. In this paper, propose a novel technique for email spam detection using SVM and feature extraction which achieves accuracy of 98% with the test datasets.

Keywords: Email Spam, Data Mining, Email Classification, Spam Detection and Prevention.

I. INTRODUCTION

The generation in the growth of data from recent decades is expanding enormously. Different sources like commercial sites, engineering field, Facebook and other social links like twitter, you-tube contributes in the size also, many-sided quality of data. To deal with and to separate importance among the data different apparatus and methods are utilized. Data Mining is a process utilized for extricating concealed and obscure information from the databases for looking for knowledge. Data can shift in measure, many-sided quality to structure. Data can be as audio, video or just a content data. To deal with and to separate the attractive properties from the data, mining is done.

II. KNOWLEDGE DISCOVERY PROCESS

Knowledge form data can be accomplished by experiencing different steps as said below. Data mining term is likewise marked as Knowledge discovery method, which implies a strategy of removing helpful information from an arrangement of raw data. Data mining is a piece of knowledge discovery as in [3], [4].

1) *Collection of Raw Data:* Dataset can be gathered from different sources like on the web and disconnected, social media sources, banks, retail division and so on.
2) *Data Selection:* Relevant data that is useful is chosen for examination.
3) *Data Pre-Processing:* Cleansing of the data to evacuate any kind of clamor, false or missing an incentive from the data is completed.
4) *Transformation:* The data is changed into proper shape with the goal that mining task can be completed.
5) *Data Mining:* Extraction of important examples from the data by utilizing different data mining strategies.
6) *Evaluation:* Extracted designs are dissected for reality value of the examples and its importance.
7) *Knowledge:* The above system mines the pertinent knowledge from the raw dataset. Knowledge can be represented by different procedures.

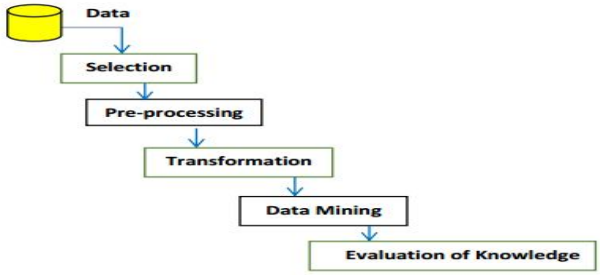

Fig. 1: Knowledge Discovery Process for mining data as in [3].

Different Techniques of Data Mining

The different procedures followed in data mining are delegated specified beneath as in [3]. The accompanying steps are executed on unessential data to pick up and get to pertinent information.

A. Anomaly Detection: Data information that is false or unimportant is mined. Anomaly detection identifies the information with no actualities.

B. Association Rule Mining (ARM): It is a process of distinguishing relationship among the attributes present in the datasets.

C. Clustering: It is a process that gatherings the comparative data in a single cluster without utilizing any predefined demonstrate. Clustering characterizes its own model and is spellbinding system of collection the data.

D. Classification: It is a process that has a predefined demonstrate and generalizes the data into known predetermined classes. Classification is a predictive model.

E. Summarization: A process of representing the data in a precise shape for perception.

III. EMAIL SPAM

Internet has turned into a vital and fundamental part of human life. The expansion in the use of internet has expanded the quantity of record holders over different social sites. Email is the least difficult and quickest method of correspondence over the internet that is utilized both by and by and professionally. Because of the expansion in the number of record holders and increment in the rate of transmission of emails a difficult issue of spam emails had stirred. From a review it was dissected that more than 294 billion emails are sent and got each day. More than 90% emails are accounted for to be spam emails as in [5].

Emails are named into two classes Spam emails and Ham emails. Spam emails are the garbage sends got from illegitimate clients that may contain promotion, malicious code, infection or to increase individual benefit from the user. Spam can be transmitted from any source like Web, Text messages, Fax and so forth., contingent on the mode of transmission spam can be classified into different classifications like email spam, web spam, content spam, social networking spam as in [1].

The rate at which email spamming is spreading is expanding colossally due to quick and forward method for sharing information. It was accounted for that client gets more spam sends than ham sends. Spam filtration is essential since spam waste time, energy, bandwidth, and storage and devour different assets as in [2]. Email can be sorted as a spam email on the off chance that it demonstrates following attributes:

1) *Unsolicited Email:* Email got from obscure contact or illegitimate contact.
2) *Bulk Mailing:* The kind of email which is sent in bulk to numerous clients.
3) *Nameless Mails:* The kind of sends in which the character of the client isn't appeared or is covered up.

Spamming is a noteworthy issue and causes genuine loss of bandwidth and cost billon of dollars to the administration suppliers. It is basic for recognizing the spam mail and ham mail. Numerous calculations are up until now used to effectively portray the sends on their conduct but since of the changing advances programmers are ending up shrewder. So better calculations with high exactness are required that effectively name a mail as spam or ham mail. Spam channel method is utilized to mark the mail as a garbage and undesirable mail and prevents it from entering the confirmed record holder's inbox. Channels can be gathered in two classifications as in [2]:

1) *Machine Learning Based Technique:* These strategies are Support Vector Machine, Multi-Layer Perceptron, Naïve Bayes Algorithm, and Decision Tree Based and so on.
2) *Non-Machine Learning Based Technique:* These strategies are mark based, heuristic filtering, dark and whitelist, sandboxing, mail header filtering and so forth.

The achievement proportion of machine learning calculations over non-machine learning calculations is more. These strategies work by choosing the best features from the data to assemble the emails as spam or ham. Feature selection can be done in two ways:

1) *Header Based Selection:* Selecting the best feature from the header of the mail. It contains sender's address, BCC (Blind Carbon Copy), CC (Carbon Copy), To, From, Date and Subject.
2) *Content Based Selection:* Selecting the best feature from the content in the mail. It contains the fundamental message either as content, audio or video, connections and so forth.

Content Based Feature Selection is demonstrated as the most verified feature selection when contrasted with Header Based as Header Based Feature Selection can be effortlessly tempered by the programmers or spammers.

IV. LITERATURE SURVEY

Izzat Alsmadi et al. 2015 [6], Information clients depend vigorously on messages' framework as one of the real wellsprings of correspondence. Its significance and utilization are consistently developing in spite of the advancement of versatile applications, informal communities, and so on. Messages are utilized on both the individual and expert levels. They can be considered as official records in correspondence among clients. Messages' information mining and examination can be directed for a few purposes, for example, Spam location and arrangement, subject characterization, and so forth. In this paper, an extensive arrangement of individual messages is utilized with the end goal of envelope and subject orders. Calculations are produced to perform bunching and order for this vast content gathering. Order in light of NGram is appeared to be the best for such huge content accumulation particularly as content is Bi-dialect (i.e. with English and Arabic substance).

A.K. Sharma et al. 2015 [7], The nonstop development of email clients has brought about the expanding of spontaneous messages otherwise called Spam. In current, server side and customer side hostile to spam channels are presented for identifying diverse highlights of spam messages. Nonetheless, as of late spammers presented some powerful traps comprising of inserting spam substance into computerized picture, pdf and doc as connection which can make ineffectual to current procedures that depends on examination advanced content in the body and subject fields of email.

Idris I et al. 2015 [8], The expanded idea of email spam with the utilization of urge mailing instruments incite the requirement for locator age to counter the danger of unsolocited email. Locator age roused by the human insusceptible framework executes molecule swarm streamlining (PSO) to produce identifier in negative determination calculation (NSA). Exception indicators are remarkable highlights created by nearby anomaly factor (LOF). The neighborhood exception factor is executed as wellness capacity to decide the nearby best (Pbest) of every applicant identifier. Speed and position of molecule swarm improvement is utilized to help the development and new molecule position of every exception locator. The molecule swarm advancement (PSO) is actualized to enhance indicator age in negative choice calculation instead of the arbitrary age of identifiers. The model is called swarm negative determination calculation (SNSA). The trial result demonstrate that the proposed SNSA display performs superior to the standard NSA. NADIR OMER et al. 2014 [9], pam messages are considered as a genuine infringement of protection. What's more, it has turned out to be exorbitant and undesirable correspondence. Despite the fact that, Support Vector Machine (SVM) has been broadly utilized as a part of email spam discovery, yet the issue of managing tremendous information is time and memory devouring and low precision. This investigation accelerates the computational time of SVM classifiers by decreasing the quantity of help vectors. This is finished by the K-implies SVM (KSVM) calculation proposed in this work. Moreover, this paper proposes a system for email spam identification in view of half breed of SVM and K-implies grouping and requires one more information parameter to be resolved: the quantity of bunches. The analysis of the proposed instrument was completed utilizing spambase standard dataset to assess the plausibility of the proposed technique. Megha Rathi et al. 2013 [10], As web is extending step by step and individuals for the most part depend on web for correspondence so messages are the quickest method to send data starting with one place then onto the next. Presently a day's every one of the exchanges all the correspondence whether general or of business occurring through messages. Email is a viable apparatus for correspondence as it spares a great deal of time and cost. Be that as it may, messages are likewise influenced by assaults which incorporate Spam Mails. Spam is the utilization of electronic informing frameworks to send mass information. Spam is flooding the Internet with many duplicates of a similar message, trying to constrain the message on individuals who might not generally get it.

V. METHODOLOGY

In this area, we will talk about the proposed methodology for email spam detection procedure. The fig. 1. Demonstrates the workflow.

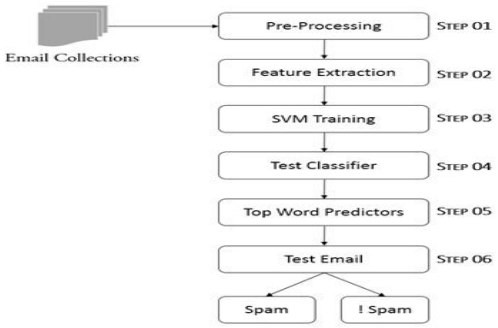

Fig. 1: Workflow for Email Spam Detection

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

A. Pre-processing

The pre-processing step is utilized to expel the noises from the email which are irrelevant and require not to be available. The pre-processing step incorporates

1) Removal of Numbers
2) Removal of Special Symbol
3) Removal of URLS
4) Stripping HTML
5) Word Stemming

B. Feature Extraction

Feature Extraction is utilized to separate the essential and important features from the email body. The feature transforms the email into 2D 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 utilized for the training necessity. The training dataset include content of spam and classifier are prepared utilizing it. Subsequent to training, the classifier is prepared to classify the spam emails.

D. Test Classifier

The classifier is tested with various training information to test the accuracy of the classifier. The proposed arrangement accomplishes up to 98 % accuracy in classifying emails.

E. Test Email

After the training stage is finished, an example email is given as input to the classifier to characterize the email. The classifier produces output in the forms of 0 or 1, 1 implies it is spam and 0 implies it is not a spam.

VI. RESULTS

In this area, detail result is clarified with each stage output. MATLAB is utilized for executing algorithm. Figures presents output of each stage as given below.

A. Preprocessing

```
==== 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. 2: Pre-processing Step

B. Feature Extraction

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

Fig. 3: Feature Extraction Step

C. Training

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

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

Fig. 4. Training Step

D. Training and Test Case Accuracy:

```
Training Accuracy: 99.850000

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

Fig. 5. Training and Test Case Accuracy

E. Email Spam or !Spam

```
==== 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.6.. Email Classification

Finally organize given an sample email, and it is classified as spam or !spam in light of its content.

VII. CONCLUSION

The issue imposed by spam emails is obvious. So, automatic prevention or filtering of spam emails is necessary for users and Internet Service Providers. Despite the fact that the features utilized as a part of email classification broadly change among various approaches in the literature survey, classification with a small set of discriminative features is favored in perspective of processing complexity. 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] J. W. Yoon, H. Kim, and J. H. Huh, “Hybrid spam filtering for mobile communication,” computers & security, vol. 29, no. 4, pp. 446–459, 2010.
[2] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[3] S. Ruggieri, “Efficient c4. 5 [classification algorithm],” IEEE transactions on knowledge and data engineering, vol. 14, no. 2, pp. 438–444, 2002.
[4] B. Sch Ikopf, S. Mika, C. Burges et al., “Input space versus feature space in kernel-based method,” IEEE Trans Neural Networks, pp. 1000–1017.
[5] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” Machine learning, vol. 29, no. 2-3, pp. 131–163, 1997.
[6] Izzat Alsmadi, Ikdam Alhami, Clustering and classification of email contents, Journal of King Saud University - Computer and Information Sciences, Volume 27, Issue 1, 2015, Pages 46-57, ISSN 1319-1578
[7] A. K. Sharma and R. Yadav, "Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Technique," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, 2015, pp. 1089-1093.
[8] Idris I, Selamat A, A Swarm Negative Selection Algorithm for Email Spam Detection. J Comput Eng Inf Technol, 2015, 4:1. doi:10.4172/2324-9307.100012
[9] NADIR OMER FADL ELSSIED, THMAN IBRAHIM, WAHEEB ABU-ULBEH, AN IMPROVED OF SPAM E-MAIL CLASSIFICATION MECHANISM USING K-MEANS CLUSTERING", Journal of Theoretical and Applied Information Technology 28th February 2014. Vol. 60 No.3
[10] Megha Rathi, Vikas Pareek, "Spam Mail Detection through Data Mining – A Comparative Performance Analysis", I.J. Modern Education and Computer Science, 2013, 12, 31-39.