International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

STUDY OF VARIOUS TECHNIQUES TO FILTER SPAM EMAILS

Jagdeep Kaur[1], Priyanka[2]

[1, 2] Computer *Science and Engineering Swami Vivekanand Institute of Engineering and Technology*

Abstract - *Email is the most proficient and quickest method of correspondence to trade data over the web. Because of the expansion in the quantity of record holders over the different social locales, there is a colossal increment in the rate of spreading of spam messages. In spite of having different apparatuses accessible still, there are many hotspots for the spam to start. Electronic spam is the most troublesome Internet wonder testing huge worldwide organizations, counting AOL, Google, Yahoo and Microsoft. Spam causes different issues that may, thus, cause financial misfortunes. Spam causes activity issues and bottlenecks that point of confinement memory space, registering power and speed. Spam makes clients invest energy expelling it. Different techniques have been produced to channel spam, including black list/white list, Bayesian classification algorithms, keyword matching, header information processing, investigation of spam-sending factors and investigation of received mails. In this paper we have discussed about spam emails and various filtering and data mining techniques to filter spam emails.*

***Key Words*: Spam, Bayesian classification algorithm, keyword matching, header information processing, Black List, White List**

1. INTRODUCTION

Internet has become an important and essential part of human life. The increase in the utilization of internet has increased the number of account holders over various social sites. Email is the simplest and fastest mode of communication over the internet that is used both personally and professionally. Due to the increase in the number of account holders and an increase in the rate of transmission of emails a serious issue of spam emails had aroused. From a survey, it was analysed that over 294 billion emails are sent and received every day. Over 90% emails are reported to be spam emails [1]. Emails are labelled into two categories Spam emails and Ham emails. Spam emails are the junk emails received from illegitimate users that might contain advertisement, malicious code, Virus or to gain personal profit from the user. Spam can be transmitted from any source like Web, Text messages, Fax etc., depending upon the mode of transmission spam can be categorised into various categories like email spam, web spam, text spam, social networking spam [2]. The rate at which email spamming is spreading is increasing tremendously because of the fast and immodest way of sharing information. It was reported that user receives more spam emails than ham emails. Spam filtration is important because spam waste time, energy, bandwidth, storage and consume other resources [3].

Email can be categorised as a spam email if it shows following characteristics [4]:

- **Unsolicited Email:** Email received from unknown contact or illegitimate contact.
- **Bulk Mailing:** The type of email which is sent in bulk to many users.
- **Nameless Mails:** The type of emails in which the identity of the user is not shown or is hidden.

Spamming is a major issue and causes serious loss of bandwidth and cost billon of dollars to the service providers. It is essential for distinguishing between the spam mail and ham mail. Many algorithms are so far used to successfully characterise the emails on their behaviour but because of the changing technologies, hackers are becoming more intelligent. So, better algorithms with high accuracy are needed that successfully label an email as spam or ham Email. Spam filter technique is used to label the email as a junk and unwanted email and prevents it from entering the authenticated account holder's inbox.

2. FILTER TECHNIQUES

Filter techniques can be grouped into two categories [3]:

1. **Machine Learning Based Technique:** These techniques are Support Vector Machine, Multi-Layer Perceptron, Naïve Bayes Algorithm, Decision Tree Based etc.

2. **Non-Machine Learning Based Technique:** These techniques are signature based, heuristic scanning, blacklist/whitelist, sandboxing and mail header scanning etc.

    - **Signatures:** Signatures contains the information taken from the documents. Signatures detect the spam or threats by generating a unique value called a hash value for each spam message. Signatures can be generated in two ways firstly by fragmenting the words into pairs and secondly by random generation of numbers. Signature uses the hash value with the

new email value to compare and to analyze if the email is spam or ham.

- **Blacklist and Whitelist***:* A blacklist is a list of spammers or any illegitimate contact that tries to send a spam or malicious email while whitelist is a list that contains legitimate users or contacts that are known to an individual account holder.
- **Heuristic Scanning:** This technique uses rules to detect malicious contents and threats. Heuristic scanning is a faster and efficient technique that detects the spam or threats without executing the file and works by understanding the behaviour. Heuristic scanning allows the user to change the rules.
- **Mail Header Checking:** In this technique set of rules are developed that are matched with the email header to detect if the email is spam or ham. If the header of the email matches the rules, then it invokes the server and directs the emails that contain empty field of "From", confliction in "To", confliction in "Subject" etc.

3. DATA MINING

Data mining, which is also defined as a knowledge discovery process, means a process of extraction of unknown and potentially useful information (such as rules of knowledge, constraints, regulations) from data in databases [7]. Data Mining follows three main steps in preparing the data for processing, reducing the data to concise it and extracting useful information. The major algorithms followed in data mining are classified into six classes [5]. The following steps are executed on raw data to obtain relevant information.

1. **Anomaly Detection:** It is the identification of data records that are not desirable and might contain an error in it, say temperature is 45, this indicates a bogus data without units.
2. **Association Rule Mining (ARM):** It is a process of identifying linkage among the items present in the database. ARM induces the relationship between the items, say bread and butter or bread and jam.
3. **Clustering:** A descriptive process that groups the data of same structure in one cluster without using a pre-defined structure say, a mail is a spam or ham mail. Clustering will group the set of data into two clusters based on the characteristics generated viz. a mail can be spam depending upon the type of content in the mail or a mail can be ham mail. Such as K-Means and K-Medoid.
4. **Classification:** A predictive process that generalizes the known structure to new data. Such as Support vector machine, Multi-Layer Perceptron.

Summarization: A process of representing the data in the compact form for visualization.

Various data mining techniques and systems are available to mine the data depending upon the knowledge to be acquired, depending upon the techniques and depending upon the databases [1].

1. **Based on techniques:** Data mining techniques comprises of query-driven mining, knowledge mining, data-driven mining, statistical mining, pattern based mining, text mining and interactive data mining.
2. **Based on the database:** Several databases are available that are used for mining the useful patterns, such as a spatial database, multimedia database, relational database, transactional database, and web database.
3. **Based on knowledge:** The knowledge discovery process, include association rule mining, classification, clustering, and regression. Knowledge can be grouped into multilevel knowledge, primitive knowledge and general level knowledge.

4. LITERATURE REVIEW

Tiago A. Almeida and Akebo Yamakami (2010) performed a comparative analysis using content-based filtering for spam. This paper discussed seven different modified versions of Naïve Bayes Classifier and compared those results with Linear Support Vector Machine on six different open and large datasets. The results demonstrated that SVM, Boolean NB and Basic NB are the best algorithms for spam detection. However, SVM executed the accuracy rate higher than 90% for almost all the datasets utilized [5].

Loredana Firte, Camelia Lemnaru and Rodica Potolea (2010) performed a comparative analysis on spam detection filter using KNN Algorithm and Resampling approach. This paper makes use of the K-NN algorithm for classification of spam emails on the predefined dataset using feature's selected from the content and emails properties. Resampling of the datasets to appropriate set and positive distribution was carried out to make the algorithm efficient for feature selection [2].

Ms.D.Karthika Renuka, Dr. T. Hamsapriya, et. al. (2011) performed a comparative analysis of spam classification based on supervised learning using several machine learning techniques. In this analysis, the comparison was done using three different machine learning classification algorithms viz. Naïve Bayes, J48 and Multilayer perceptron (MLP) classifier. Results demonstrated high accuracy for MLP but high time consumption. While Naïve Bayes accuracy was low than MLP but was fast enough in execution and learning. The accuracy of Naïve Bayes was enhanced using FBL feature selection and used filtered Bayesian Learning with Naïve Bayes. The modified Naïve Bayes showed the accuracy of 91% [6].

Rushdi Shams and Robert E. Mercer (2013) performed a comparative analysis of the classification of spam emails by using text and readability features. This paper proposed an efficient spam classification method along with feature selection using the content of emails and readability. This paper used four datasets such as CSDMC2010, Spam Assassin, Ling Spam, and Enron-spam. Features are categorized into three categories i.e. traditional features, test features and readability features. The proposed approach is able to classify emails of any language because the features are kept independent of the languages. This paper used five classification based algorithms for spam detection viz. Random Forest (RF), Bagging, Adaboostm 1, Support Vector Machine (SVM) and Naïve Bayes (NB). Results comparison among different classifiers predicted Bagging algorithm to be the best for spam detection [7].

Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta and Anuja Arora (2014) performed a comparative analysis of text and images by using KNN, Naïve Bayes and Reverse-DBSCAN Algorithm for email spam detection. This analysis paper proposed a methodology for detecting text and spam emails. They used Naïve Bayes, K-NN and a modified Reverse DBSCAN (Density- Based Spatial Clustering of Application with Noise) algorithm. Authors used Enron dataset for text and image spam classification. They used Google's open source library, Tesseract for extracting words from images. Results show that these three machine learning algorithms give better results without pre-processing among which Naïve Bayes algorithm is highly accurate than other algorithms [8].

Savita Pundalik Teli and Santosh Kumar Biradar (2014) performed an analysis of effective email classification for spam and non-spam emails. In this paper, the author compares three classification techniques such as KNN, Support Vector Machine and Naïve Bayes. She shows that Naïve Bayes gives maximum accuracy among other algorithms that is 94.2%. The author then proposed a method to enhance the efficiency of Naïve Bayes. The proposed method is divided into three phases. In first phase the user creates rule for classification, second phase trains the classifier with training set by extracting the tokens, and in third phase based on maximum token matches, the email is classified as spam or ham. The performance of Naïve Bayes is improved by this Algorithm. They take parameters Precision, Recall, Accuracy [9].

Izzat Alsmadi and Ikdam Alhami (2015) performed an analysis on clustering and classification of email contents for the detection of spam. This paper collected a large dataset of personal emails for the spam detection of emails based on folder and subject classification. Supervised approach viz. classification alongside unsupervised approach viz. clustering was performed on the personal dataset. This paper used SVM classification algorithm for classifying the data obtained from K-means clustering algorithm. This paper performed three types of classification viz. without removing stop words, removing stop words and using N-gram based classification. The results clearly illustrated that N-gram based classification for spam detection is the best approach for large and Bi-language text [10].

Ryan McConville, X. Cao, W. Liu, P. Millerv (2016) gives a general framework to accelerate existing algorithms to cluster large-scale datasets which contain large numbers of attributes, items, and clusters is proposed. This framework makes use of locality sensitive hashing to significantly reduce the cluster search space. This framework has a guaranteed error bound in terms of the clustering quality. This framework can be applied to a set of centroid-based clustering algorithms that assign an object to the most similar cluster. K-Modes categorical clustering algorithm to present how the framework can be applied is adopted. The framework with five synthetic datasets and a real world Yahoo! Answers dataset was used. The experimental results demonstrate that this framework is able to speed up the existing clustering algorithm between factors of 2 and 6, while maintaining comparable cluster purity [11].

N. Akhtar and N. Agarwal (2014) introduces a new process for segmenting a photograph. It combines two learning algorithms, particularly the k-approach Clustering and Neutrosophic good judgment, together to obtain effective results through eliminating the uncertainty of the pixels. Forming hard clusters by applying neutrosophy before k – mean algorithm [12].

C. Jacob and K A Abdul Nazeer (2014) made use of k-means algorithm along with improved clustering process ant colony algorithm. The algorithm works on the principle of probability following pick and drop. The algorithm is capable enough for determining optimal number of clusters and their corresponding centroid's. It eliminates the problem of local optimal hence making the algorithm least dependent on initial centroid's [13].

5. PROPOSED WORK

The existing methods have some limitations of having less accuracy and precision. The main objective of this proposed work is to upgrade the existing machine learning techniques in distinguishing spam emails.

Objectives of proposed work are as follows:

1. To study various spam detection algorithms for emails.
2. To propose an approach for email spam detection using improved MLP with N-gram feature selection.
3. To compare and analyse the results of proposed approach with the existing on the basis of parameters

viz. Accuracy, Sensitivity, Specificity, Root Mean Square Error and Precision.

6. CONCLUSION

In this paper, we have discussed about spam emails and its characteristics. Further we elaborate various filtering techniques like machine learning based technique and non-machine learning based technique in which we cover signatures, balcklist and whitelist, heuristic scanning, mail header checking. We discussed about data mining, its procedure and its techniques. Literature survey discusses about the previous work. Existing system has some drawbacks related to accuracy and precision. So to enhance the current machine learning techniques, we proposed the work using MLP with N-gram feature selection.

REFERENCES

[1] D. Jurafsky and J. H. Martin, "N-Gram," Speech and Language Processing, 2014.

[2] L. Firte, C. Lemnaru, and R. Potolea,"Spam Detection Filter using KNN Algorithm and Resampling", in 6th International Conference on Intelligent Computer Communication and Processing -IEEE, pp.27-33, 2010.

[3] Kaur, R. K. Gurm, "A Survey on Classification Techniques in Internet Environment", International Journal of Advance Research in Computer and Communication Engineering (IJARCCE), vol. 5, no. 3, pp. 589–593, March 2016.

[4] H Kaur, P. Verma, "Survey on E-Mail Spam Detection Using Supervised Approach with Feature Selection," International Journal of Engineering Sciences and Research Technology (IJESRT), vol. 6, no. 4, pp. 120-128, April 2017.

[5] B. Yu and Z. Xu,"A comparative study for content-based dynamic spam classification using four machine learning algorithms", Knowledge Based System-Elsevier, vol. 21, pp. 355–362, 2008.

[6] D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques", in 2011 International Conference on Process Automation, Control and Computing - IEEE, pp. 1–7, 2011.

[7] R. Shams and R. E. Mercer, "Classifying spam emails using text and readability features", in International Conference on Data Mining (ICDM) - IEEE, pp. 657–666, 2013.

[8] A. Harisinghaney, A. Dixit, S. Gupta, and Anuja Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm", in International Conference on Reliability, Optimization and Information Technology (ICROIT)-IEEE, pp.153-155, 2014.

[9] S. P. Teli and S. K. Biradar, "Effective Email Classification for Spam and Non- spam", International Journal of Advanced Research in Computer and software Engineering, vol. 4, 2014.

[10] Alsmadi and I. Alhami, "Clustering and classification of email contents", Journal of King Saud University - Computer and Information Science -Elsevier, vol. 27, no. 1, pp. 46–57, 2015.

[11] R. M. Conville, X. Cao, W. Liu, P. Millerv, "Accelerating Large Scale Centroid-Based Clustering with Locality Sensitive Hashing", in- International Conference on Data Engineering- IEEE, pp. 649-660, 2016.

[12] N. Akhtar, N. Agarwal , "K-Mean Algorithm For Image Segmentation Using Neutrosophy",in- International Conference on Advances in Computing Communications and Informatics- IEEE, pp. 2417-2421, 2014.

[13] C. Jacob, K A Abdul Nazeer, "An Improved ICPACA based K- Means Algorithm with Self Determined Centroids, "in- International Conference on Data Science and Engineering- IEEE , pp. 89-93 , 2014.