Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVII, Issue II, July-Dec 2023

A Review: Environmental Awareness

[1]Naveen Kumar Kedia, [2]Dr. U.K. Pareek ,[3]Aditya Jaiswal, [4]Abhay Sharma

[1] Assistant Professor, Department of Information Technology, JECRC College
[2]Professor, Department of Mathematics, JECRC College
[3]B.Tech Student, Department of Information Technology, JECRC College
[4]B.Tech Student, Department of Information Technology, JECRC College

ABSTRACT

Nowadays, a big part of people rely on available email or messages sent by the stranger. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests. and income, by taking behavioural aspects into consideration makes these methods an efficient one as compares to others. Steals useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising. Spam email is a kind of commercial advertising which is economically viable because email could be a very cost effective medium for sender.

Keywords - Term Frequency, Inverse Document Frequency, language tool kit.

[1] INTRODUCTION

In recent years, internet has become an integral part of life. With increased use of internet, numbers of email users are increasing day by day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. Email has now become one of the best ways for advertisements due to which spam emails are, so customer segmentation is becoming very popular and also became the efficient solution for this existing problem. Spam is a huge waste of everybody's time and can quickly become very frustrating if you receive large amounts of it. Identifying these

spammers and the spam content is a laborious task. Spam emails, also known as non-self, are unsolicited commercial or malicious emails, sent to affect either a single individual or a corporation or a bunch of people.

[2] RELATED WORK

A solution is proposed as distinguish the customers group into two groups named as premium and standard with the help of machine learning methods named as NEM, LiRM and LoRM.

Tushar Kansal, Suraj Bahuguna, Vishal Singh, Tanupriya Choudhury. "Customer Segmentation using K-means Clustering", International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS).2018, In this paper customer segmentation on Telecom customers is achieved by using information such as age, interest, etc. with the help of cluster analysis method.

Use Case Diagram

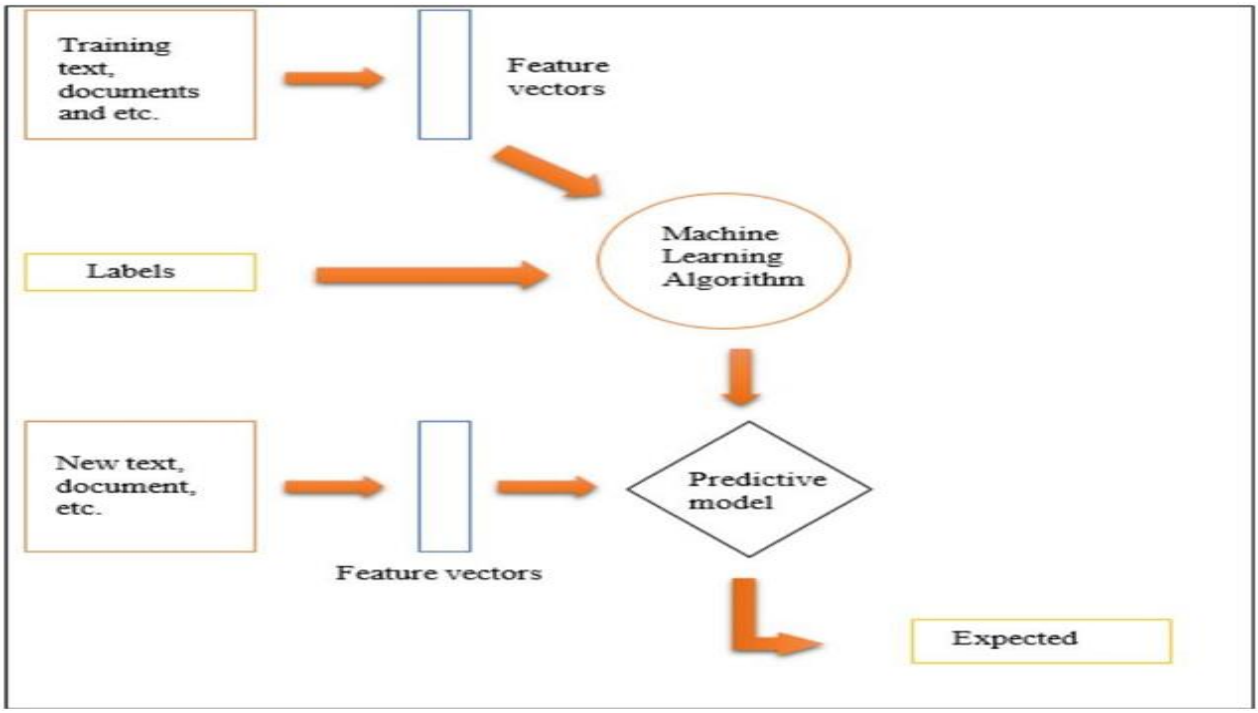

Figure 1. Use case Diagram

Authors described about cyber attacks phishers and malicious attackers are frequently using email services to send false kinds of messages by which target user can lose their money and social reputations. These results into gaining personal credentials such as credit card number, passwords and some confidential data .In this paper, authors have used Bayesian Classifiers .Consider every single word in the mail. Constantly adapts to new forms of spam.

The machine learning model will be a pre-trained model with a feedback mechanism to distinguish between a proper output and an ambiguous output. This method provides an alternative architecture by which a spam filter can be implemented.

Proposed system attempts to use machine learning techniques to detect a pattern of repetitive keywords which are classified as spam. The system also proposes the classification of emails based on other various parameters contained in their structure such as Cc/Bcc, domain and header.

They observed that Bi-gram algorithm performs best in spam detection in both datasets.

A new approach based on the strategy that how frequently words are repeated was used. The key sentences, those with the keywords, of the incoming emails have to be tagged and thereafter the grammatical roles of the entire words in the sentence need to be determined, finally they will be put together in a vector in order to take the similarity between received emails.

[3] Proposed System

In this system, to solve the problem of spam, the spam classification system is created to identify spam and non- spam. Since spammers may send spam messages many times, it is difficult to identify it every time manually .So we will be using some of the strategies in our proposed system to detect the spam. The proposed solution not only identifies the spam word but also identifies the IP address of the system through which the spam message is sent so that next time when the spam message is sent from the same system our proposed system directly identifies it as blacklisted based on the IP address.

In the proposed model ,the web application is done using dot net and spam detection is done using machine learning .The web application consists of following modules:

User Management:
The user who is using this for the very first time must register, by using the website the user or the individual should get registered into it, by registering this will help to maintain separate account for each user. Registration of the user is must before they log in. The user will login to the main page with his registered name and password. Once the user successfully login the authorized page will be displayed otherwise that shows the error messages. Login is compulsory.

Login:

The user will login to the main page with his registered name and password. Once the user successfully login the authorized page will be displayed otherwise that shows the error messages. Login is compulsory.

Registration:

First time while using the website the user or the individual should get registered into it, by

Figure 2 spam detection

Registering this will help to maintain separate account for each user. Registration of the user is must before they log in.

Compose

Input: the sender will compose the new email; the

sender should add the address of the recipient, the subject andthe message.

Output: the email will be sent based to the addressmentioned by the recipient.

Inbox

Figure 3. API Key

This page will store all of the mails received by user. Allthe received Mails will be listed sorted in order of date.

Input: the inbox page will accept all the incoming emails sentto an individual.

Output: the receiver can open and read the email received totheir address.

Sent

This folder stores all the mails sent from the user.

Trash

This folder will store all of mails deleted by the user.Input: and Delete all the unwanted emails.

Output: all the deleted emails are added in the trash bin.Trash bin stores all the deleted emails.

Voice Message

Input: The Email has been sent in the form of the textmessage by the sender

Output: The email has been read through the use of voicenote by the receiver.

Offline notification

Input: The sender sends an email

Output: the receivers receive a notification offline in the textformat as SMS.

Delete For everyone

Input: here the sender deletes the email which he has sent Output: the email has been erased or deleted for both thesender as well as the receiver.

Read Message

Input: The receiver will read the email.

Output: the sender will get a notification stating the sender as read the message.

When we receive message in the inbox ,that message will be exported to dataset. This message will be detected as spam or not using Naïve Bayes Classifier.

Before detecting whether received message is spam or not

,the model has to be trained which is explained in the below section.

4. Implementation/Deployment:

All the implementation process of the spam detection by using machine learning based binary classifier project will be presented.

Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVII, Issue II, July-Dec 2023

After spam detection ML model has been trained, an Api key will be generated by the server in order to deploy to the web service. Figure 2. API generated by ML studio server

Then, the api will be entered into API key from the VW algorithm using visual studio code. Figure 3. Insert API key into form in VS After the API key has been confirmed by the server, the spam ML web service will be shown below:

Figure 4 Spam ML web service

Word Search Classifier then created to test the accuracy of detection by using VW algorithm which works by dividing messages into bigrams.

```
public void Train(string trainingFile)
{
    bigrams = new Dictionary<string, int>();
    hamCutover = 0;
    string[] trainingData = File.ReadAllLines(trainingFile);
    List<LabeledMessage> messages = new List<LabeledMessage>();

    // read messages
    foreach (string line in trainingData)
    {
        if (line.StartsWith("Ham") || line.StartsWith("Spam"))
        {
            string[] data = line.Split(new char[] { ',' }, 2);
            messages.Add(new LabeledMessage(data[0], data[1]));
        }
    }
```

Figure 5 Messages divide into bigrams

The result of detection will be counted using the formula as below:

```
public ClassifierValidationResult(List<LabeledMessage> messages, TimeSpan elapsedTime)
{
    LabeledMessages = messages;
    Correct = messages.Count(x => x.ModelClassification == x.RealClassification);
    Total = messages.Count;
    Wrong = Total - Correct;
    Accuracy = (Decimal)Correct / Total;
    ElapsedTime = elapsedTime;
}
}
```

Figure 6 Scoring result formula

5. Results

In this section, the result for spam probability, time elapsed and comparison of spam detection using different malware detection is presented. This section presented the results based on experiments and study of this project. The entire graph above focused on the comparison of detection using different malware detection tools which is Joe Sandbox cloud, hybrid analysis (Falcon sandbox) and visual studio.

Joe Sandbox is used for hardware virtualization to analyse and detect malware. For malware analysis, it intercepts execution, extracts additional information and returns/continues execution. For this project, Joe Sandbox view is used to view detection status by constructing own project query.

This tool let user to use any of 1500 identifiers and comparison operators. In this case, operators such as special characters and repetitive text are chosen to filters the data. Joe Sandbox will provide full analysis report that contains ID, sample name, MD5 and etc.

6. Conclusion

The performance of a classification technique is affected by the quality of data source. Irrelevant and redundant features of data not only increase the elapse time, but also may reduce the accuracy of detection. It also able to score the model and weight them successfully.

During the implementation, only text (messages) can be classified and score instead of domain name and email address.

This project only focus on filtering, analysing and classifying message and do not blocking them.

Hence, the proposed methodology may be adopted to overcome the flaws of the existing spam detection.

7. REFERENCES

[1] Anitha, PU & Rao, Chakunta & , T.Sireesha. (2013). A Survey On: E-mail Spam Messages and Bayesian Approach for Spam Filtering. International Journal of Advanced Engineering and Global Technology (IJAEGT). 1. 124- 136.

[2] Attenberg, J., Weinberger, K., Dasgupta, A., Smola, A., & Zinkevich, M. (2009, July). Collaborative email-spam filtering with the hashing trick. In Proceedings of the Sixth Conference on Email and Anti-Spam.

[3] Awad, W. A., & ELseuofi, S. M. (2011). Machine learning methods for spam e-mail classification. International Journal of Computer Science & Information Technology (IJCSIT), 3(1), 173-184.

[4] Barnes, J. (2015). Azure Machine Learning. Microsoft Azure Essentials. 1st ed, Microsoft.

[5] Chang, M. W., Yih, W. T., & Meek, C. (2008, August). Partitioned logistic regression for spam filtering. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 97- 105). ACM.

[6] Çıltık, A., & Güngör, T. (2008). Time-efficient spam e-mail filtering using ngram models. Pattern Recognition Letters, 29(1), 19-33. 48