International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1404-1407 www.ijaem.net ISSN: 2395-5252

Detection of Anomalies in Email Traffic Using Naïve Bayes

Prof. Sheela Jawale , Sagar Jadhav, Kirti Ingle, Nidhi Shetty
Department of Information Technology Datta Meghe College of Engineering, Airoli Navi Mumbai, India
Department of Information Technology Datta Meghe College of Engineering, Airoli Navi Mumbai, India
Department of Information Technology Datta Meghe College of Engineering, Airoli Navi Mumbai, India
Department of Information Technology Datta Meghe College of Engineering, Airoli Navi Mumbai, India

Submitted: 01-06-2021 Revised: 14-06-2021 Accepted: 16-06-2021

ABSTRACT— E-mail spam is a very recent problem for every individual. The e-mail spam is nothing but an advertisement of any company/product or any kind of virus which is receiving by the email client mailbox without any notification. To solve this problem the different spam filtering technique is used. The spam filtering techniques are used to protect our mailbox from these spam mails. In this project, we are using the Naïve Bayesian Classifier for emails classification. The Naïve Bayesian Classifier is very simple and efficient method for classification of ham and spam mails. Here we are using the dataset for classification of spam and non-spam mails. The feature extraction technique is used to extract the feature. The result is to increase the accuracy of the system.

Keywords—Naive bayes,feature extraction,spam mail,ham mail;

I. INTRODUCTION

Email spam detection has become more important topic now a days due to increasing phishing attacks.Email spam targets singular clients with regular postal mail messages. Email spam records are regularly made by checking Usenet postings, taking Internet mailing records, or scanning the Web for locations. .In this System the received mail will go in the preprocessing model where the Counter Vectorization will take place .The work of the Counter Vectorization is to convert the text into number's as machine's understands number's better than text messages. In this process it will break the messages into separate Token format.The number of times each Token occurs in the message is counted by the Counter Vectorizer .Once the Preprocessing is done it will compare the number of occurrence count with the dataset containing ham and spam statements .After that Classifying Model will extract and test the features using Naïve Bayes Classifier which will classify the mail into ham and spam mail.Lastly with the help of timestamp the spam mail will get deleted automatically .

II. FINDINGS FROM LITERATURE SURVEY

- From the Literature survey it is found that Unsupervised Machine Learing algorithms are not best suited for Spam detection as it gives poor prediction results.
- K nearest neighbor is a simple algorithm that stores the available cases and classifies the new cases on the basis of similarity measure.
- KNN has been used in statistical estimation and pattern recognition which is useful for our system but it doesn't work well with data of high dimensions. Hence it doesn't work well with the large data set, in our case it doesn't work well with long list of emails.
- Local Outlier Factor (LOF) is an unsupervised machine learning algorithm that can be used for anomaly detection. It computes the local density deviation of a given data point with respect to its neighbors.
- It considers the outliers as the samples that has lower density than their neighbors. So scalability is a major issue in LOF as the computations are memory sensitive
-

III. AIM & OBJECTIVE

The necessity of the system is to provide a solution for the incoming spam mails in an individual's system.The basic aim of the system is to classify the incoming mails as ham and spam using naive bayes classifier and further deleting the spam mail from the system.The objective of this system is to discover the pattern of data ,nmake prediction based on previous pattern of data,automatically detect the anomalies i.e spam mails in the system anf further deleting the detected spam mails from the system.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1404-1407 www.ijaem.net ISSN: 2395-5252

IV. PROPOSED SYSTEM

The proposed system uses naive bayes classifier for the detection of incoming anomalies i.e spam mails .The system consists of a registration page for the users to get registered in the system through which the users can access the further components of the system.The system further consists of a login page for entering the system and provides the user with a compose box for sending mails to a destination emailid.A dataset containing ham and spam messages has been provided to the system for classification purposes.Naive bayes has been used in system for classification purposes.The system consists of an inbox listing the incoming spam and ham messages.

SOFTWARE AND PACKAGES REQUIRED IN PROPOSED SYSTEM:

SOFTWARE REQUIREMENTS:

1. PYTHON 3.8.2
2. ANACONDA 3
3. SPYDER
4. SQL YOG

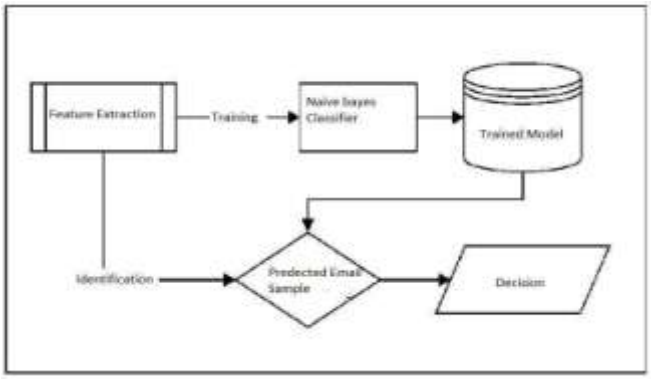
PACKAGE REQUIREMENTS:

1. SKLEARN
2. PICKLE
3. COUNTERVECTORIZER
4. MULTINOMIAL NB

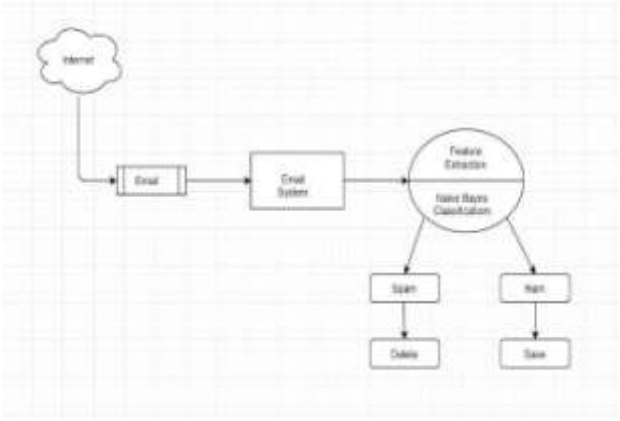
V. METHODOLOGY

Acquire a dataser containing ham and spam mails respectively and perform feature extraction on the dataset using different packages.After feature extraction the extracted dataset is then divided into training data and test data.70 percent of data is used for training data and rest 30 percent is used as test data.The training data set is given to the naive bayes classifier for development of a training model based on the classification of the naive bayes classifier.The new mail which will enter the system through the means of user logged into the system will be compared with the trained model for the classification purpose.The internal process will contain the tfidvectorizer and countervectorizer to assign and count the number of times the word has appeared in the mail.The system will compare the email sample and trained model and then give a decision.The ham mail will get saved and spam mail will get deleted from the system using time stamp.

VI. ARCHITECTURE



International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1404-1407 www.ijaem.net ISSN: 2395-5252



DFD Diagram

VII.IMPLEMENTATION



Fig.1 Registration Page



Fig.2 Login Page



Fig.4 Result Page



Fig.3 Compose Box



Fig.5 Inbox List

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1404-1407 www.ijaem.net ISSN: 2395-5252

VIII. WORKING OF PROPOSED SYSTEM

1. Initially the user registers on the registration page with name,username,phone number and password.
2. After registration the user will have to login through the login page to enter into the system.
3. After successful login,the user will get a compose box containing sender mail id,receiver mail id and compose message through which the user can send a message.
4. After the user sends a message,the classifier will classify the message and show the message as ham or spam depending on the message.
5. The inbox will then list all the messages containing ham and spam mails.
6. In The system will automatically save the ham message and delete the spam message according to the time stamp assigned to it.

IX. ANALYSIS AND CONCLUSION

This proposed system indicates that the supervised learning algorithm naive bayes is simple to understand and execute.The proposed system using this classifier can be processed without complexity and can be trained easily.This System determines the accuracy and classification of mails without any hassle.A larger dataset can be efficiently trained and tested with the system.According to the analysis,a same number of ham and spam mails would be more efficient in training the model.The accuracy of the system can be increased by increasing the layers of the system.

The system provides security to the user by assigning username and passwords through which the user can login and safeguards the system by automatically deleting the spam mails by using timestamp feature.The system provides flexibility as it can be installed on different platforms and can be easily modified with advanced features.

REFERENCES

[1] Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana," Relevance Feature Discovery for Text Mining"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 27, NO. 6, JUNE

[2] Feldman, Moshe Fresko,Yakkov Kinar ,"Text Mining at the Term Level", Ronen Feldman , Moshe Fresko , Yakkov Kinar , Yehuda Lindell , Orly Liphstat , Martin Rajman , Yonatan Schler , Oren Zamir

[3] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2010, pp. 753–76

[4] 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, Vol. 10