SPAM MAIL DETECTION USING DECISION TREE

Priya Pandey
School of Computer Science and Engineering Galgotia's University Greater Noida,Uttar Pradesh, India
anushkaguptably@gmail.com

Govil
School of Computer Science and Engineering Galgotia's University Greater Noida,Uttar Pradesh, India
govilthakur7037@gmail.com

Rohit Singh
School of Computer Science and Engineering Galgotia'sUniversity Greater Noida,Uttar Pradesh, India
rs318301@gmail.com

Dr. Anupam Sharma
Associate Professor,
School of Computer Science and Engineering
Galgotia's University
Greater Noida, Uttar Pradesh, India
anupam.sharma@galgotiasuniversity.edu.in

Abstract—
The most annoying issue on the Internet, electronic spam impacts multi-national firms. Spam can lead to the number of issue including financial loss. Spam creates bottlenecks and congestion that decrease speed, processing power, and huge storage. People took time to eliminate spam. To filter spam, many methods have been developed, such blacklists/whitelists, Bayesianclassification methods, keyword matching, header data processing, spam source analysis, and incoming email analysis. In this study, three multi-layer perceptron design machine learning approaches are described for accurately and efficiently separating spam from email messages. The decision tree classification, multi-layer auto - encoders, and naive Bayes classification are samples of commonly used techniques. All of them are utilized to train data on genuine or spam emails. The results are then discussed, and the consequences of the methods considered are matched to the proposed models Email has become the main form of communication in recent years and is often the target of active or passive attacks. To combat such operations, early establishment of good spam filtering systems is vital.

Currently, there are numerous effective spam filters with a range of performance, and its average accuracy range between 60 to 80%. The most of filtering technique, however, are not able to handle the rapidly moving situations than develop while spam emails are steadily handed to spammers. As a result, the logistic regression tree classifier and the decision tree encoder are examined at and assessed. The Outstanding Requirement is an improved spam control mechanism or a boost in the capacity of several existing data mining techniques. email spam is filtered. According to basis of empirical evidence, LMT has the highest performance levels and about 90% accuracy in recognizing among junk mail and non-spam emails.

I. INTRODUCTION:-

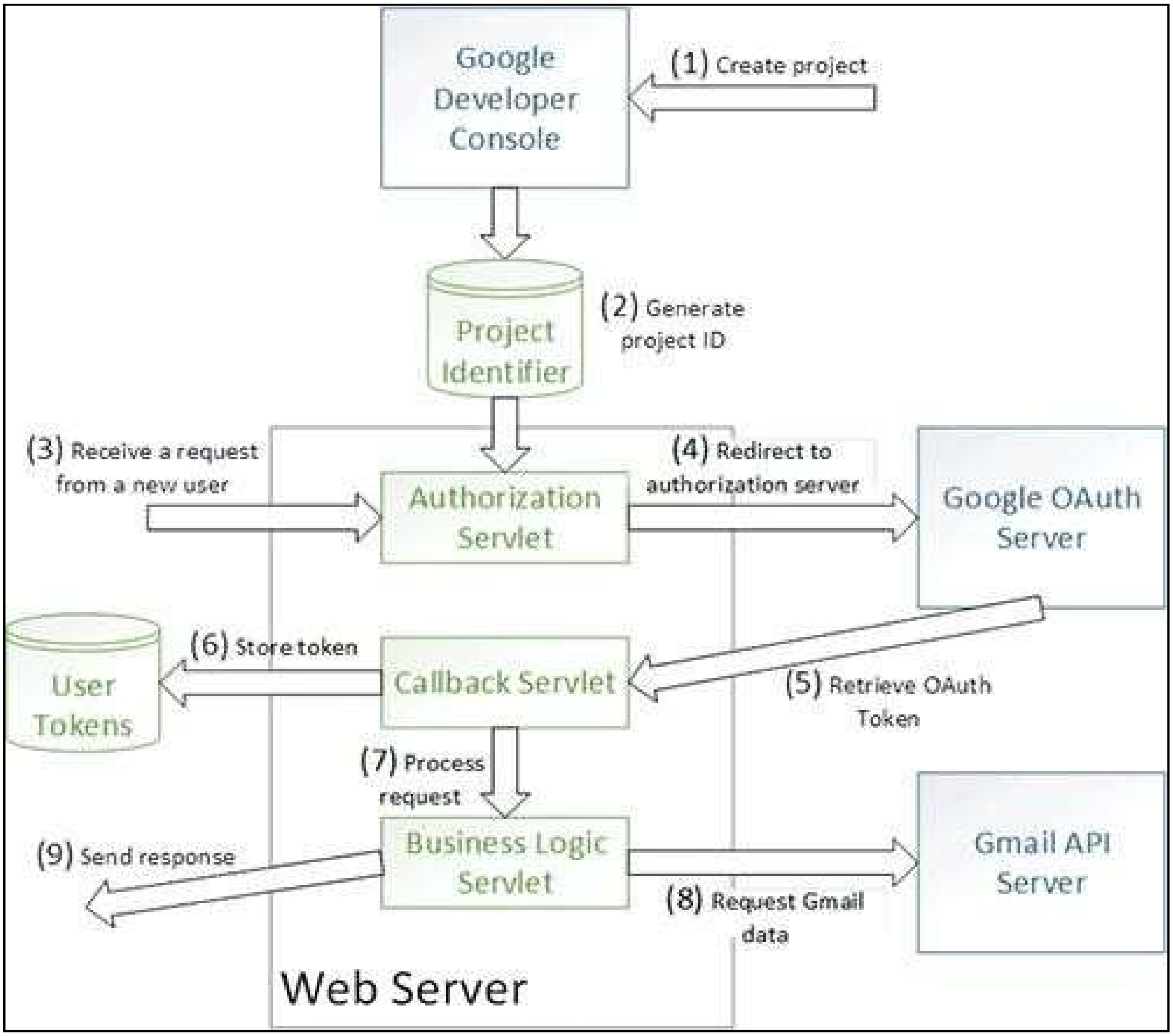
An effective way of share information is via email. The use of email and the expansion of the Internet also caused concern about just the rise of spam. On the World Wide Web, spam can originate from anyone. Even with anti-spam solutions, its prevalence is growing daily.Assessing the tools at the availability of companies that can even measure the amount of spam is one method to gauge the present situation. This involves corporate email systems, gateways, spam filters, and end- user learning. In the entireInternet, people cannot avoid this critical issue. Without a computerized anti-spam system, the internet would be filled with spam, harming Internet access and resulting in a lowest amount of bandwidth. A whitelist isa collection of email addresses that consumers regularly receive email from. Apart domain inputs & functional domains, users can also enter email addresses. The benefits of whitelisting is that users or administrators may indicate their preferred contacts so that legitimate mail received from whitelisted addresses is not reported as spam when mail is received from other communicators. a functionality which thus lets you add e-mail addresses to a list.

A blacklist is a collection of email addresses that consumers typically do not receive messages from. A collection of rules that are applied as follows are used in the email header verification process: The email is identified as spam and dispatched to the spam folder if the header corresponds with a training data header that has already been blacklisted. It will be approved if not. Many anti-spam techniques are depend on Bayesian detection. Nowadays, only the more efficient text algorithms are used for filtering. Historically, various rule-based software packages have been used for filtering operations. Rule-based solutions have two major drawbacks. First, these systems required users to create a set of rules. Users needed extensive knowledge of spam to formulate appropriate rules. Second, these rules had to be reconstructed by experts as the nature of spam changed over time.Basically, the reformulation takes longer and has a higher error rate. The quantity of information stored at the place of digital file and database has grown at staggering rate over pastdecade. On the same period, user of this data expect to search ever most sophisticated information and the data patterns hidden therein. Marketing managers are no longer satisfied with just contacting customers, they also need detailed information about their customers, including less exhaustive information. Information mining procedures and technique have been develop to meet the need. Today, eighty percentage of information is stored in text form — magazine, newspaper, document, e-mail, etc. To extract hidden information from such text data, many formats are utilised. Text mining, also known as information mining, is employed. Text mining is a knowledge-intensive process in which people interact with document collections over time while making use of a variety of analytical tools.

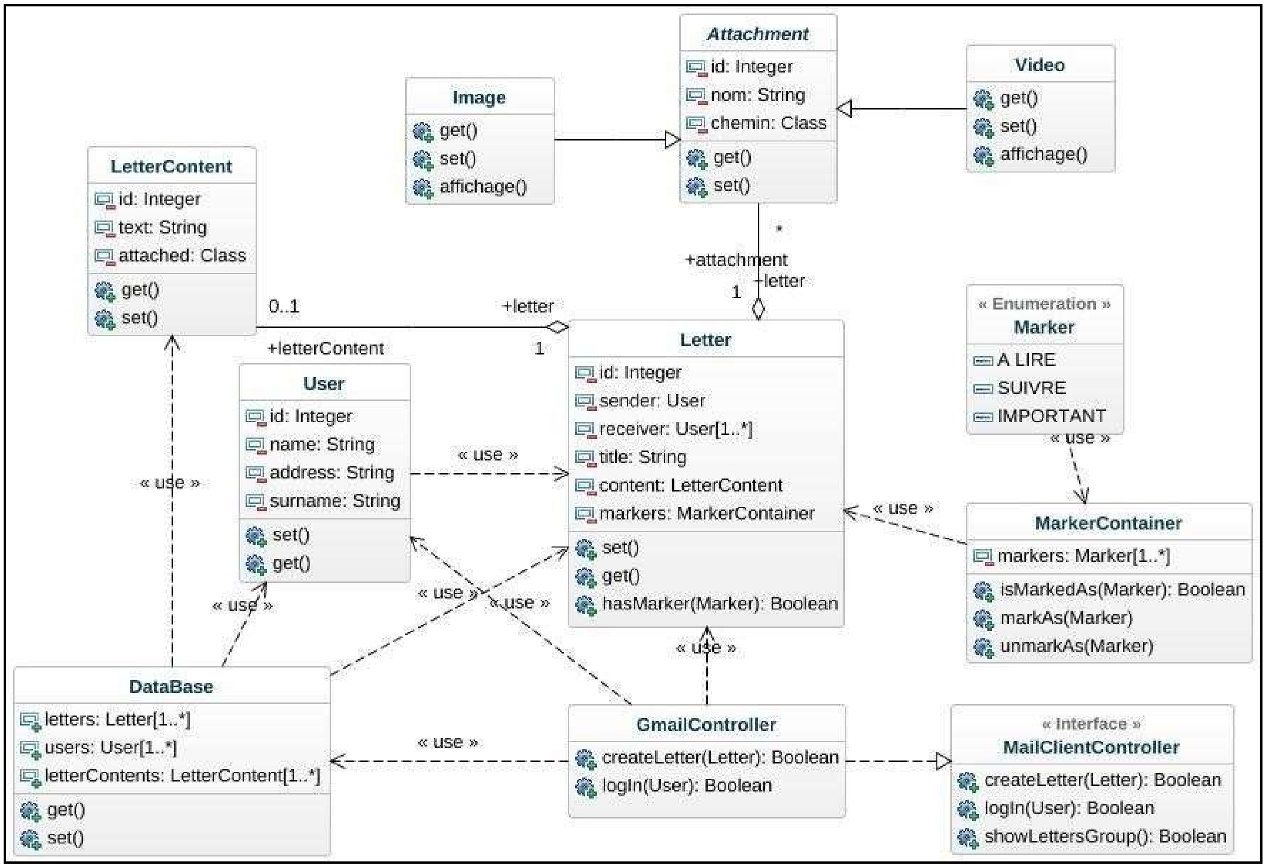
LITERATURE SURVEY:-

In the age of information technology, the exchange of information has become very easy and fast. Users can connect with her people around the world on various platforms. Email is the easiest, most affordable and fastest way to disseminate information on a global scale. Due to its simplicity, email is also vulnerable to a variety of attacks, spam being the most prevalent and destructive. No one wants to receive email that has nothing to do with them because it wastes time and resources. Additionally, these emails may contain malicious content hidden as attachments or URLs that can compromise the security of the host system. Spam is an unrelated and unsolicited message oremail sent by an attacker to a large number of recipients using information sharing email or other media. As a result, there is a great need for security in email systems. Spam emails may contain Trojanhorses, rats, and viruses. Attackers primarily use this strategy to trick people into using Internet services. They may send spam emails with multiple attachments and URLs stuffed with malicious spam websites, leading to identity theft, financial fraud, and invasion of privacy. There is a possibility. Many email service providers allow their customers to create her keyword-based email filtering rules. However, this is difficult and users don't want to customize their emails, so spammers are targeting users' emails as a result of this approach

DATA FLOW DIAGRAM

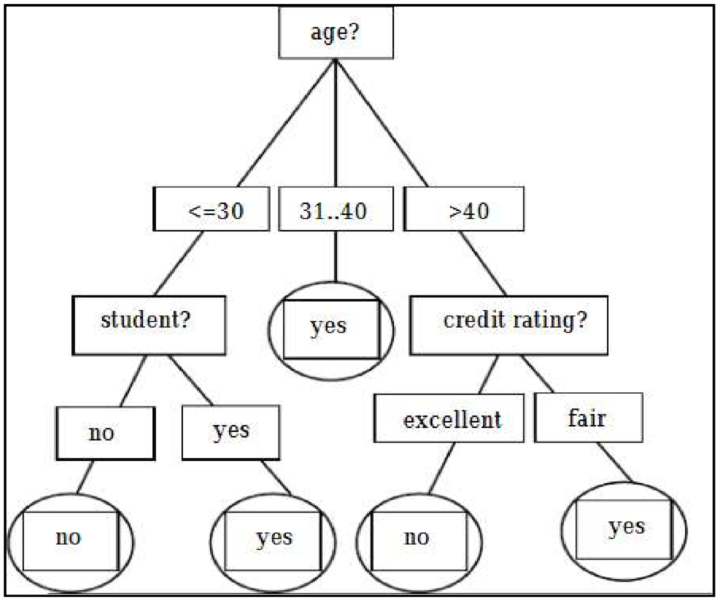


UML DIAGRAM:



METHODOLOGY/IMPLEMENTATION:-

In contrast to previous data mining classifiers, this paper uses a range of decision tree classifiers but concentrates especially on decision tree classifiers for certain purposes. in spam filtering engineering. This is due to the simplicity of implementation and comprehension of decision tree filters. Overall, it provides adequate performance in terms of spam detection. A method used frequently in data mining is decision tree learning. The objective is to build a DT model and train it to forecast a target variable's value based on a variety of input variables. The example that follows. One of the input variables corresponds to each internal node. There is a child advantage for each conceivable value of this input variable. Given the value of the input variable, each leaf indicates the value of the target variable, which is represented by the path from root to leaf. By dividing the source sentences into several subsets according to the values of key attributes, the tree can be "learned". Each derived subset is subjected to this process being repeated recursively. An example of this is a recursive partition. When a subset of nodes all share the same value for the target variable or when the split adds no more values to the prediction, recursion is finished.



Whenever we want to handle e-mail, we requires Mail.jar file and Activation.jar file but using of these file increases probability of runtime error to avoid this errors we use maven project and gives dependency and uses group-Id as com.sun.mail and it gives required jar file to our project, because we want java mail API and all classes that is present in mail API thats why we requires dependency. There are some important classes present in API that is javax.mail.messaging Exception, javax.mail.Session and javax.mail.message used for send message, subclass of The transport.send() method is used to send the message when it is created, its characteristics and content are filled out, and it is sent. All of the applications use decision tree to filter mail and sometimes it detects wrong however organizations are working on it to fix it but it still detecting some mails as spam and actually they are not spam. People are facing problems because of that so, we are working on that. So, we are going to make anapplication in which we are going to show all mails on notification section.

There are2 section on notification are one is mail and another one is for spam mail. Through which user can easily see all mail which are spam or which are not spam. As we know Google provides us APIs for use their google services like google map, Gmail, etc. In this Application we are using Gmail API to access Gmail account of the user and Gmail API provides OAuth 2.0 protocol to authentication google account and authorizing access to user data, for using this API we have to create an android application that can be creates using android studio and logical programming, for managing mails we will use java programming language. There are some libraries used in this application one of them is volley library, used for speeds up networking for android apps while while making it simpler. Volley enables network request scheduling that is automatic. We will use a machine learning algorithm to filter mail.

RESULT AND ANALYSIS AND CONCLUSION:-

After implementing the decision tree method for the email function corpus results, we evaluated its performance and selected the best one. The scoring is based on the three-weighted averages of Precision, Recall, and F-Measure. According to the definition of accuracy, it is "the percentage of items found that are relevant items measured as the ratio of relevant items searched to the total number of items searched." "The proportion of related items retrieved as determined by the relationship between the number of related items retrieved and the total number of related items in the collection," is how recall is defined. The definition of recall, the third component of the F measure, is "2 * Precision * Recall / (Precision + Recall), is a combined measure of precision and recall." Table 1 shows the average precision, recall rate, and F- value score for each decision tree approach. Furthermore, Figure shows how significantly different the weighted averages of these decision tree techniques are.

In conclusion, research that evaluated the efficacy of six decision tree algorithms for classifying emails found that the Decision Stump approach is the best classification algorithm for spam emails.. The algorithms were applied to a record corpus and their results were compared. -Utala Malaysia's email service is called Mails for University. Additionally, without compromising classification accuracy, the technique for applying document counts to email features resulted in areduction in the number of features in email records. We know that computing the Better characteristics that can be used in the classification process are produced by the tf-IDF weighting approach on the decreased features. The study also showed that they recommended the best algorithms for addressing spam issues in email systems. Regarding his advice for this study, the suggested solution calls for the elimination of voice stopwords during the preparation of the email data, making it suitable for use in various email systems around the world. I can manage it. Future research in this field will look into alternative preprocessing.

REFERENCES

[1] T. Verma, N. S. J. I. J. o. I. T. Gill, and E. Engineering, "Email Spams via Text Mining using Machine LearningTechniques," 9, no. 4, pp. 2535-2539, (2020).

[2] N. Saidani, K. Adi, M. S. J. C. Allili, and Security, "A semantic-based classification approach for an enhanced spam detection," 94, p. 101716, (2020)