Data Mining for Prediction and Classification of Engineering Students achievements using Improved Naïve Bayes

Ms. Ashna Sethi[1], Mr. Charanjit Singh[2]
[1](Department of Computer Science and Engineering , RIMT-IET , Mandi Gobindgarh) [2](Assistant Professor, Department of Computer Science and Engineering , RIMT-IET , Mandi Gobindgarh)

Abstract— : Data mining is one of the emerging trends in the field of research. Educational Data mining is one of the upcoming trends to rule over the industry. EDM helps the educational firms to make decisions related to the student's performance. Now a days, with a higher education dropping out of students' has been increasing, it affects not only the students' career but also on the reputation of the institute. The proposed system makes use of the Improved Naive Bayesian mining technique for the extraction of useful information. The experiment is conducted on 300 engineering students' of Gulzar Group of Institutes with 21 attributes. Result proves that Improved Naïve Bayesian algorithm provides more accuracy over other methods like Regression, Decision Tree etc., for comparison and prediction. The system aims at increasing the success graph of students using Improved Naïve Bayesian and the system which maintains all student admission details, course details, subject details, student marks details, attendance details, etc. It takes student's academic history as input and gives students' upcoming performances on the basis of semester. The comparison between Naïve Bayes and Improved Naïve Bayes is made using parameters like: FP Rate, Precision, Recall, TP Rate, F-meausre, Kappa Statistics and Accuracy. The Result shows that improved Naïve Bayes has given better performance then only Naïve Bayes. The accuracy of Improved Naïve Bayes is 86.66 %.

Index Terms— Data Mining, EDM, Naïve Bayes, Decision tree.

I. INTRODUCTION

The data mining is "The process of extracting comprehensible, unknown, and actionable information from large databases and using it to make important business decisions" Data mining is concerned with the analysis of data for finding hidden and unexpected pattern and relationships in large volume of data. Basically the focus of data mining is to find the information which is hidden and unexpected and convert it into the understandable form for future use. Data mining is also called as KDD, knowledge discovery in databases.

Data mining techniques are listed below:

1. Classification: Is used to place the data in predetermined group.
2. Clusters: Data items are placed in a group according to logical relationships.
3. Associations: Data mining is applied to data set to find out the associations.
4. Sequential Patterns: Data is mined to expected behaviour patterns and trends.

Knowledge discovery steps are:

- Data cleaning:- to remove noise , irrelevant and inconsistent data from the database
- Data integration:- where multiple data sources may be combined to build a data set
- Data selection:- where the data relevant to the analysis are selected from the data base
- Data transformation:- the data are transformed into the form appropriate for mining
- Data mining: - the process where intelligent methods are applied in order to extract data patterns from data set.
- Pattern evaluation: - evaluate patterns to identify the truly interesting patterns representing knowledge.

1. Knowledge representation: - where knowledge representation techniques are used to present the minded knowledge.

Educational Data Mining: The educational mining is used to investigate students' gaining knowledge of conduct. The intention of the have a look at is to expose how beneficial facts mining can be utilized in higher education to beautify students' common performance. He used students' facts from database direction and accrued all available records collectively with educational data of students, direction records. Then done the superior records mining (IDM) strategies to discover many forms of information along with centroid primarily based absolutely, distribution based and density based totally clustering. Cluster consists of businesses with small distance a few of the cluster individuals. Also this will have clustered the scholar into organizations the usage of Centroid, and detected all

similarities within the facts mining assessment. Finally, this will show how we're capable of experience the located know-how to enhance the general performance of pupil. The information mining strategies, specially categorized to help in enhancing the splendid of the higher academic device by using comparing scholar information to have a look at the number one attributes that might have an impact on the pupil normal overall performance in guides. The extracted class suggestions are based on the unique information the extracted class policies are studied and evaluated.

There are distinctive varieties of academic environment are:

- Off-Line Education (Traditional Class room): It is used to deliver expertise and competencies based totally on face-to-face touch. In this we need to recollect college students 'conduct, overall performance, curriculum, and so on. That became accrued in study room surroundings.
- E-Learning and Learning Management System (LMS): ELearning gives online guidance. LMS also offers conversation, collaboration, administration and reporting tools. Web Mining (WM) techniques had been applied to students 'information saved via those structures in log documents and databases.
- Intelligent Tutoring System (ITS) and Adaptive Educational Hypermedia System (AEHS): It adapts teaching to the desires of every unique scholar. DM has been implemented to facts picked up via these structures, together with log files, user models, and many others.

Predicting overall performance, those students are on chance may be find out and some remedial action may be taken to improve their performance. Due to this results of the University can improve. The educational facts mining can be used to get the remarks for the teachers so the instructor can enhance teaching method. Also, the records may be useful for people who are designing the path contents. There are many facts mining techniques may be used for this like class, clustering, regression, rule mining.

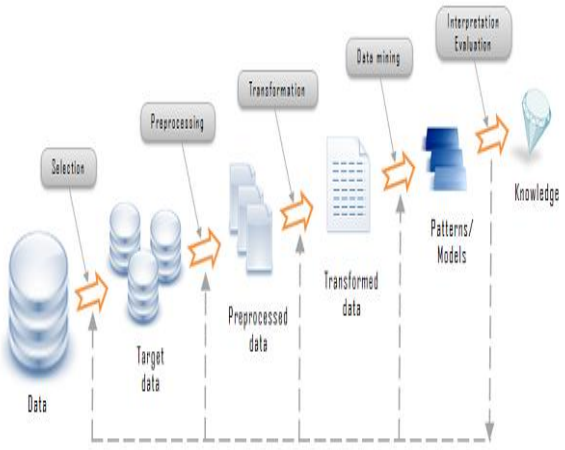

Fig.1 Knowledge Discovery.

II. RELATED WORK

Abaidullah et al. [3] presented the analysis of student's feedback data for decision making by educational community responsible for monitoring and reviewing the effectiveness of educational programs and for improving the quality of teaching and learning experience for their students. For this purpose k-means clustering algorithm is used.

Amornsinlaphachai. P, [1] has studied that to select a data mining model to predict learners' academic performance in computer programming subject to group learners for cooperative learning by comparing the efficiency of the models created from data mining with classification technique. To develop a model for cooperative learning via web using the selected data mining model to group learners. The efficiency of seven models created from data mining with classification technique by using seven algorithms that are Artificial Neural Network, K-Nearest Neighbor, Naive Bayes, Bayesian Belief Network, JRIP, ID3 and C4.5 is compared and it was found that the models created from C4.5 has the best efficiency.

Buniyamin et al. [2] highlights the importance of using student data to drive improvement in education planning. He describes the development of a tool that will enable faculty members to identify, predict and classify students based on academic performance measured using Cumulative Grade point average (CGPA) grades. His work elaborates a brief overview of the most commonly used classifiers techniques in educational data mining and an outline of the use of Neuro-Fuzzy classification in a case study research to predict and classify students' academic achievement in an Electrical Engineering faculty of a Malaysian public university.

La Red et al. [16] described the various mining models and discussed the main results. Mining models of clustering, classification and association were considered especially. It seeks to analyze patterns of success and failure for students in academics, therefore predicting the likelihood of dropping out or having poor academic performance of students, with the advantage of being able to do it early, allowing addressing action to reverse this situation.

Shaukat.K et al. [8] has focused on recognizing, extracting and calculating data associated to the learning method and improving student's performance. The purpose of our study is to evaluate the performance of students by taking different attributes like academic achievements (CGPA), gender, class test grade, environment of class, Fund/Scholarships/Private etc. In our research we will use classification and clustering techniques to analyze student performance. The techniques used in our work are decision tree, Bayesian classification-mean algorithms, neural networks, Naive's Bayes, Web based system and nearest neighbor methods.

Shahiria.A.M et al. [7] has highlighted on how the prediction algorithm can be used to identify the most important attributes in a student's data. We can really improve student's achievement and success more effectively in an efficient way using EDM techniques. It could bring the benefits and impacts to students, educators and academic institutions.

Devasai.T. et.al. [4] Has used compared different techniques of data mining on EDM like: Naive Bayesian, Regression, Decision Tree, Neural networks the proposed system is a web based application which makes use of the Naive Bayesian mining technique for the extraction of useful information. The experiment is conducted on 700 students' with 19 attributes in Amrita Vishwa Vidyapeetham, Mysuru. Result proves that Naïve Bayesian algorithm provides more accuracy over other methods like Regression, Decision Tree, Neural networks etc., for comparison and prediction.

Alshareef et al. [15] explained the effectiveness of data mining techniques in the context of higher education by offering a data mining model for the higher education system at Sebha University. In this research, association rules were used to evaluate students' performance by applying the apriori algorithm on survey data. In this task author extract knowledge that describes students' performance, which helps in identifying earlier trends in the choices of major and in helping new students to select their major.

Taylan et al. [5] introduced a systematic approach for the design of a fuzzy inference system based on a class of neural networks to assess the students' academic performance. Fuzzy systems have reached a recognized success in several applications to solve diverse class of problems. Currently, there is an increasing trend to expand them with learning and adaptation capabilities through combinations with other techniques. Fuzzy systems-neural networks and fuzzy systems-genetic algorithms are the most successful applications of soft computing techniques with hybrid characteristics and learning capabilities. The developed method used a fuzzy system augmented by neural networks to enhance some of its characteristics like flexibility, speed, and adaptability, which is called the adaptive neuro-fuzzy inference system (ANFIS). New trends in soft computing techniques, their applications, model development of fuzzy systems, integration, hybridization and adaptation are also introduced. The parameters set to facilitate the hybrid learning rules for the constitution of the Sugeno-type ANFIS architecture is then elaborated. The method can produce crisp numerical outcomes to predict the student's academic performance (SAP). It also provides an alternative solution to deal with imprecise data. The results of the ANFIS model are as robust as those of the statistical methods, yet they encourage a more natural way to interpret the student's outcomes.

Huang et al. [6] presented the study for the first time that develops and compares four types of different mathematical models to predict student academic performance in engineering dynamics - a high-enrollment, high-impact, and core course that many engineering undergraduates are required to take. The four models are the multiple linear regression model, the radial basis function network model, the multilayer perception network model, , and the support vector machine model. The inputs of the models include student's CGPA, grades earned in four pre-requisite courses, and scores on three dynamics mid-term exams. Based on the four mathematical models and six different predictor variable combinations , a total of 24 predictive mathematical models were developed from the present study. The analysis shows that the type of mathematical model has a minor effect on the average prediction accuracy (APA) and on the percentage of accurate predictions (PAP). The combination of predictor variables has only a slight effect on the APA, but a profound effect on the PAP. In general, the support vector machine models have the highest PAP as compared to the other three types of mathematical models. The work from the present study implies that the instructor should choose the simplest mathematical model, if the goal of the instructor is to predict the average academic performance of his/her, dynamics class as a whole. The simplest model which can be adopted is the multiple linear regression model, with student's cumulative GPA as the only predictor variable. Adding more predictor variables does not help improve the average prediction accuracy of any mathematical model. However, if the goal of the instructor is to predict the academic performance of individual students, the instructor should use the support vector machine model with the first six predictor variables as the inputs of the model, because this particular predictor combination increases the percentage of accurate predictions, and most importantly, allows sufficient time for the instructor to implement subsequent educational interventions to improve student learning.

The analysis and assessment of the students' feedback in improving the educational environment as well as enhancing students' learning experience is one of the critical issues for the higher education community. The conventional methods of analysis and assessment are not sufficient to explore the hidden information from the student feedback data repositories.

III. METHODOLOGY

Data mining has its great advantage in helping many organisations to find the hidden patterns in the organisational data to predict the behaviour of customers, products and processes. As the organisational databases size is increasing and the volume of the data being stored in databases is increasing so there is need of some techniques to summarize these data, identify true interesting trends and patterns from these databases, and act upon the outputs.

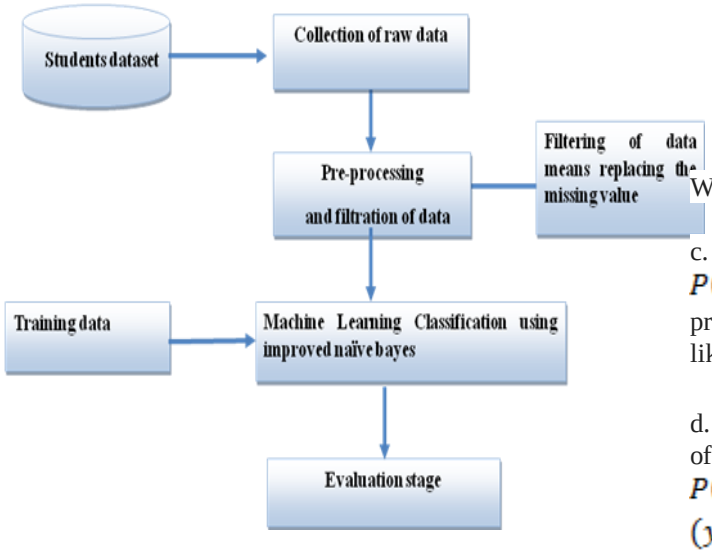

Fig.2 Steps involved in Implemented Work

A. *Dataset and Collection of Raw data*

A data of almost 300 engineering students is collected from Gulzar Group of Institutes. From the collected data, two datasets are formed namely, Training dataset and Testing dataset. Fig. 3 shows training dataset used in the work. All the values of the attributes are non-numeric. Total of 29 attributes are considered while observing the performance of the students.

B. *Preprocessing*

In preprocessing step, Noise from data is removed and also missing values are handled (either removed or replaced) using a filter.

C. *Improved Naïve Baye's Algorithm*

On Filtered data, Improved Naive Baye's algorithm is performed to predict the desired results in an efficient way following the below mentioned steps:

Step I. For each instance in dataset,

If (numeric)

Throws exception

Else

Evaluate merits of splitting the attribute into naïve bayes and decision table using forward AS search.

Step II. Find the probabilities of each word using naïve bayes as:

a. Let M be training set of tuples and their associated class labels be represented by:

$$Y= y_1, y_2, y_3 \cdots \cdots y_n \quad (1)$$

b. If a tuple Y is given,the classifier will predict that Y belongs to the class having the highest posterior probability, conditioned on Y. That is, the naive Bayesian classifier predicts that tuple Y belongs to class Ci if and only if

$$P(C_i|Y) > P(C_j|Y) \text{ for } 1 \le j \le m, j \ne I \quad (2)$$

Thus, we maximize. According to the Bayes' theorem

$$P(C_i|Y) = \frac{P(Y|C_i)P(C_i)}{P(Y)} \quad (3)$$

Where A and B are events and P(B)≠0

c. The term P(Y) is constant for all classes, only $P(Y|C_i)P(C_i)$ needs to be maximized . If the class prior probabilities are not known, then the classes are equally likely, that is,

$$P(C_1) = P(C_2) == P(C_m) \quad (4)$$

d. Reducing computation in evaluating, the naive assumption of class-conditional independencies made. Thus,

$$P(Y|C_i) = \prod_{k=1}^{n} P(y_k|C_j)$$
$$(y_1 |C_i) \times (y_2 |C_i) \times \ldots \times P((y_n|C_i). \quad (5)$$

e. Predicting the class label of Y, $P(Y|C_i)P(C_i)$ is evaluated for each class C_i,

$$P(Y|C_i)P(C_i) > P(Y|C_j)P(C_j) \quad ,1 \le j \le m, j \ne i \quad (6)$$

The predicted class label is the class C_i for which $P(Y|C_i)P(C_i)$ is the maximum.

Step III. Find the probabilities of each word using decision table as:

a. Decision table for dataset Q with n attributes, B_1, B,...,B_n, consists of :

Schema R (B_1, B_2,...,B_n, Class, Sup, Conf).

Row =(b_{1i}, b_{2i}, ..., b_{ni}, c_i,sup_i, $conf_i$) in table R represents a classification rule, where (1 ≤ j ≤ n) can be either from DOM(B_i) or a special value ANY c_i ϵ{c_1, c_2, ...,c_m}

minsup ≤ sup_i, ≤1

minconf ≤$conf_i$ ≤1

Where minsup and minconf are predetermined thresholds.

The interpretation of the rule is (A_1= a_1) and (A_2 = a_2) and …….. and (A_n = a_n) then class = c_i with probability $conf_i$ and having support sup_i,where aj≠ ANY, 1≤ j≤ n.

Finding rows whose attribute values are either ANY or equal to the corresponding attribute values of u (b_{1u}, b,, ...,b_{nu}), searching for the matches in a decision table could result in following:

i. One matching row is found:

ii. More than one matching row is found:

I. based on confidence and support:

$$C_u^k = \{C_i | Conf_i = \max_{j=1}^{k} Conf_j\} \quad (7)$$

II. based on weighted confidence and support:

$$C_u^k = \{C_i | Conf_i * Sup_i = \max_{j=1}^{k} (Conf_j * Sup_j)\} \quad (8)$$

The class of the first matching row can be simply assigned t sample to be classified.

iii. No matching row is found: The method used for classifying such samples is to use the default class.

Step IV. The overall class probability is computed as P (DT) * (P (NB)/ prior probability of class)
Assuming X (DT) = set of attributes in the DT X (NB) =set of attributes in NB

D. *Evaluation Stage*
After training the dataset, test set is evaluated based on following parameters using cross-validation method. Precision and recall, F-Measure, Accuracy. Cross validation , TP Rate, Kappa Statistics is a model used to overcome problems faced by existing models. In this method, entire data set is not used while training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on ``new'' data.

IV. EXPERIMENTAL RESULTS

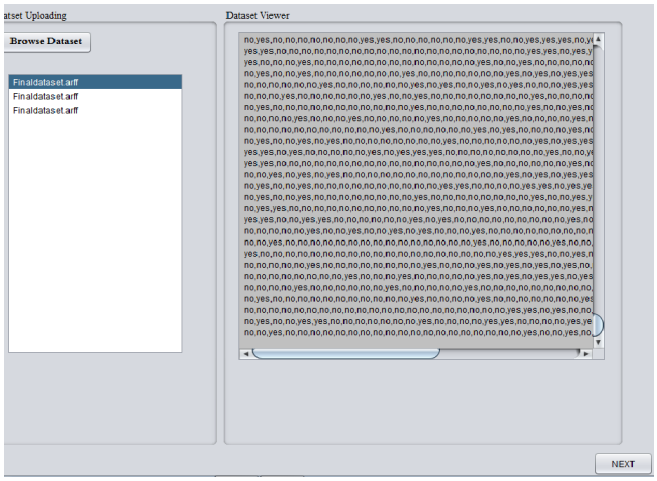

Fig 3. Date set used

Table 1: Accuracy Table

Parameter	Naïve Bayes	Improved Naïve Bayes
Accuracy	83.67%	86.67%

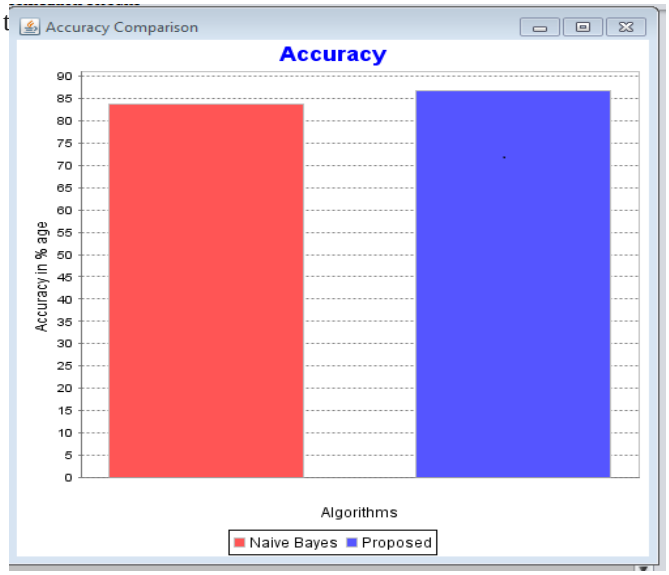

Fig 4. Accuracy Comparison Graph

Table 2: Shows comparison between Naive Bayes and Improved Naive Bayes based on TP Rate.

Parameter	Naïve Bayes	Improved Naïve Bayes
TP Rate	0.837	0.867
Precision	0.844	0.863
Recall	0.837	0.867
F-Measure	0.822	0.86

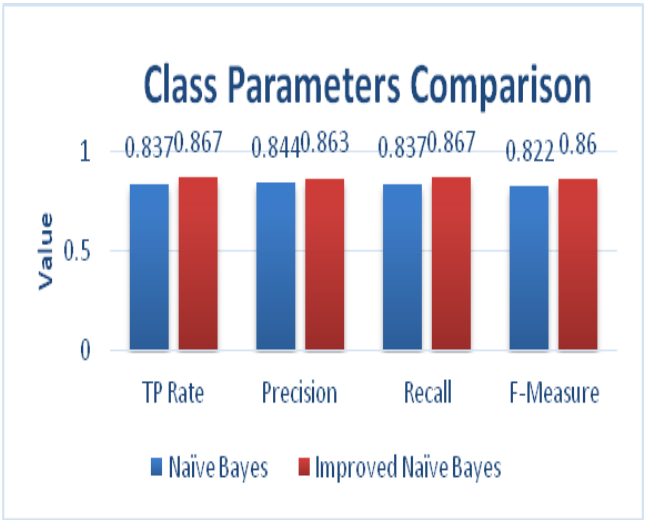

Fig 5. Class parameters Comparison Graph

Table 3: Comparison of kappa statistics value

Parameter	Naïve Bayes	Improved Naïve Bayes
Kappa Statistics	0.7433	0.7963

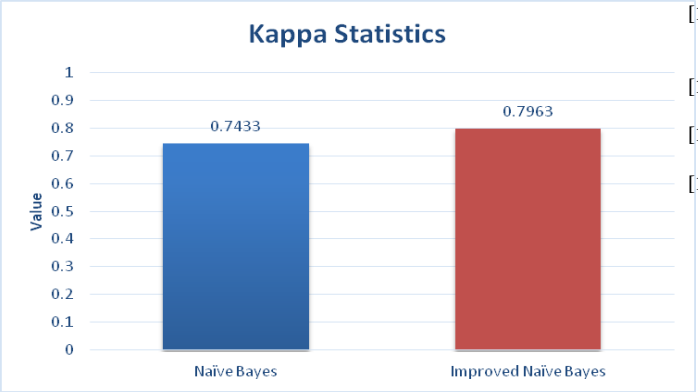

Fig 6. Kappa Statistics Comparison graph

CONCLUSION

In this paper, the classification is employed in student information to predict the students' division on the premise of previous information. As there are several approaches that area unit used for knowledge classification, Naive theorem is employed here. Information like group action, class test, seminar and assignment marks were collected from the students' previous information, to predict the performance at the top of the semester. This study can facilitate the students and the lecturers to boost
the students of all category to perform well. This study helps to spot out those students who require special attention , minimize the failure ratio and to take acceptable action for upcoming semester examination. Future work includes applying data processing techniques for Associate in nursing distended knowledge set with additional typical attributes to urge correct and economical results.

. REFERENCES

[1] Rushdi Shams and Robert E. Mercer, "Classification spam emails using text and readability features," IEEE 13th International Conference on Data Mining, 2013.
[2] Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora , "Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN Algorithm, " International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, India, Feb 6-8 2014.
[3] Masurah Mohamad and Ali Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," IEEE International Conference on Computer Communication, and Control Technology (14CT 2015), April. 2015.
[4] Izzat Alsmadi and Ikdam Alhami, "Clustering and Classification of email contents," Journal of King Saud University-Computer and Information Sciences, vol. 27, pp. 46-57, Jan. 2015.
[5] Ms.D.Karthika Renuka, Dr.T.Hamsapriya, Mr.M.Raja Chakkaravarthi, Ms.P.Lakshmisurya, "Spam Classification based on Supervised Learning using Machine Learning Techniques," IEEE, 2011.
[6] Megha Rathi and Vikas Pareek, "Spam Email Detection through Data Mining-A Comparative Performance Analysis," I.J. Modern Education and Computer Science, vol. 12, pp. 31-39, 2013, available on http://www.mecs-press.org/.
[7] Savita Pundalik Teli and Santosh Kumar Biradar, "Effective Email Classification for Spam and Non-spam," International Journal of Advanced Research in Computer and software Engineering, vol. 4, June 6, 2014.
[8] Rekha and Sandeep Negi, "A Review on Different Spam Detection Approaches," International Journal of Engineering Trends and Technology (IJETT), Vol. 1, May 6, 2014.
[9] Guanting Tang, Jian Pei, and Wo-Shun Luk,"Email Mining: Tasks, Common Techniques, and Tools", School of Computing Science, Simon Fraser University, Burnaby BC, CANADA.
[10] Seongwook Youn and Dennis McLeod," A Comparative Study for Email Classification", University of Southern California, Los Angeles, CA 90089 USA.
[11] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, "Comparative Study on Email Spam Classifier using Data Mining Techniques", IAENG.
[12] Geerthik.S," Survey on Internet Spam: Classification and Analysis", Int.J.Computer Technology & Applications,Vol 4 (3), pp. 384-391.
[13] Reena Sharma, Gurjot Kaur," Spam Detection Techniques: A Review" International Journal of Science and Research (IJSR), 2013.