ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 9, September 2016**

# Auto Clustering Emails with Naive Bayes

Prof. Suchita Walke[1], Monali Patil[2]

HOD, Dept. of Information Technology, YTCEM College, Mumbai, India[1]

M.E. Student, Dept. of Computer Engineering, SESGOIFE College, Mumbai, India[2]

**ABSTRACT:** In lot of communication e- mail plays important role. E- mail system is used for communication in all type of organizations. It is self-evident that e-mail has become a central means for the discussion of engineering work and sharing of digital assets that define the product and its production process. Engineering communication research has shown that the volume of communication is indicative of progress being made within an engineering project.
So that e-mail conversations increases as product grows and data in communication also increases. It gets difficult to handle the data at emails. So need of classification of emails. Here we have studied different classification techniques which help us to classify the large email data.

KEYWORDS**:** Emails, Clustering, Classification, Naive Bayes

## I. INTRODUCTION

Project Management is the process and activity of planning, organizing, motivating, and controlling resources, procedures and protocols to achieve specific goals in scientific or daily problems. A project is a temporary endeavor designed to produce a unique product, service or result with a defined beginning and end (usually time-constrained, and often constrained by funding or deliverables), undertaken to meet unique goals and objectives, typically to bring about beneficial change or added value. In this process lot of emails get generate. These email data need to be cluster for understanding the projects flow, reduce search time, arrange emails with label etc. There should be process which handles these things automatically. That means clustering of the emails according to projects or project category. Which help user to handle multiple project at a time. For this we are proposing artificial & NLP based technique with the use of Naive Bayes algorithm.

. The main purpose of energy efficient algorithm is to cluster emails which are coming from employee with using text categorization the algorithm is implemented here is navies bayes which calculate probability of each word and then it will be classify.

## II. RELATED WORK

In recent paper of text mining, having different type of techniques are used such as Decision tree(DT),Support vector machine(SVM),Artificial Neural Network(ANN) & K-Nearest Neighbour(KNN).from these techniques it is clear that the txt miming can be done using different mathematical formulas. In [3]the support vector m/c offers a principled approach to ML(machine learning) problem.SVM Construct their solution as weighted sum of sv which are only a subset of training input. In[1]decision tree make multistage decision's that split feature space into the associated with various class. Upon arrival of new feature vector sequence decision are made by travelling tree from top to bottom & making final decision at leaf level where class assignment rules are utilized.
In [2] it consists of input and output layers with hidden layer in between each of varying or constant number of neurons that operate as simple computing element to mimic the biological single propagation while minimizing the empirical risk during training phase.

.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 9, September 2016**

## III. PROPOSED ALGORITHM

We propose here the Naive Bayes Classifier for clustering the emails. In email text categorization Dimensionality reduction (DR) is a very important step. DR techniques can classify into Feature Extraction (FE) approaches and feature Selection (FS), as discussed below.

- Feature Extraction
- Feature Selection
- Learning Algorithm
- Classification

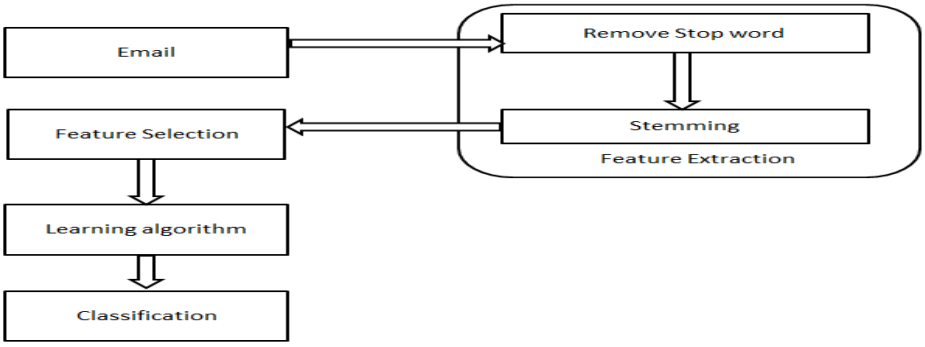


Fig.1. Shows flow of proposed algorithm

Feature Extraction: Feature Extraction is first step of preprocessing which is used to presents the text documents into clear word format. Removing stops words and Stemming words is the preprocessing tasksA document is treated as a string and then partitioned into a list of tokens. Removing stop words: Stop words such as “the”, “a”, “and”… etc are frequently occurring, so the insignificant words need to be removed. Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form eg. Connection to connect, computing to compute etc.

Feature Selection: After feature extraction the important step in pre-processing of email text classification, is feature selection to construct vector space or bag of words, which improve the scalability, efficiency and accuracy of a text classifier. The main idea of FS is to select subset of feature from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Hence feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

Learning algorithm: Implementation of proposed system will be done by following algorithms. Bayesian Theorem and network for data mining and Naive Bayes for classification.
Following are the formulas used – Baye’sFormula Mathematically it is represented as –
Let $B_1$ ,$B_2$ , $B_3$ …….$B_n$ be a partition of Ω (space) such that P($B_n$ ) ≠ 0 for any
n = 1, 2, 3… and let P (A) ≠ 0.Then,

$$P(A|B_{n,}) = \frac{P(B_n|A)P(B_n)}{\sum P(B_n|A)P(B_{n)}}$$

Classification: after calculating probability document will be classify into particular category.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 9, September 2016**

## IV. PSEUDO CODE

Step 1: Accept the input as Email document.
Step 2: Apply Feature extraction which include R classiferemoving of Stop word and steaming.
Step 3: Apply feature selection on whole document which used to text classification to reduce the dimensionality feature space and improve the efficiency and accuracy of classifier.
Step 4: Apply Learning Algorithm to document

$$\text{Posterior} = \frac{Prior\ X\ likehood}{Evidence}$$

$$P(A|B_{n,}) = \frac{P(B_n|A)P(B_n)}{\sum P(B_n|A)P(B_n)}$$

Step 5: Classify the document into particular category.

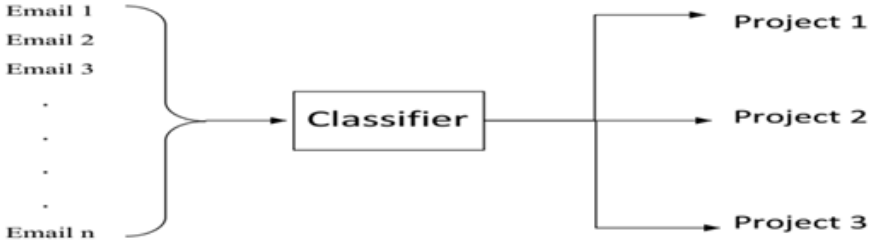


Fig.2. Shows classification of Emails

## V. SIMULATION RESULTS

The simulation studies involve the study of email classification which uses text categorization using naive bayes algorithm. The first page shows project manager login, Employee login and admin login. In this project manager can retrieved mails and employee send mails .admin part will do the training of emails.

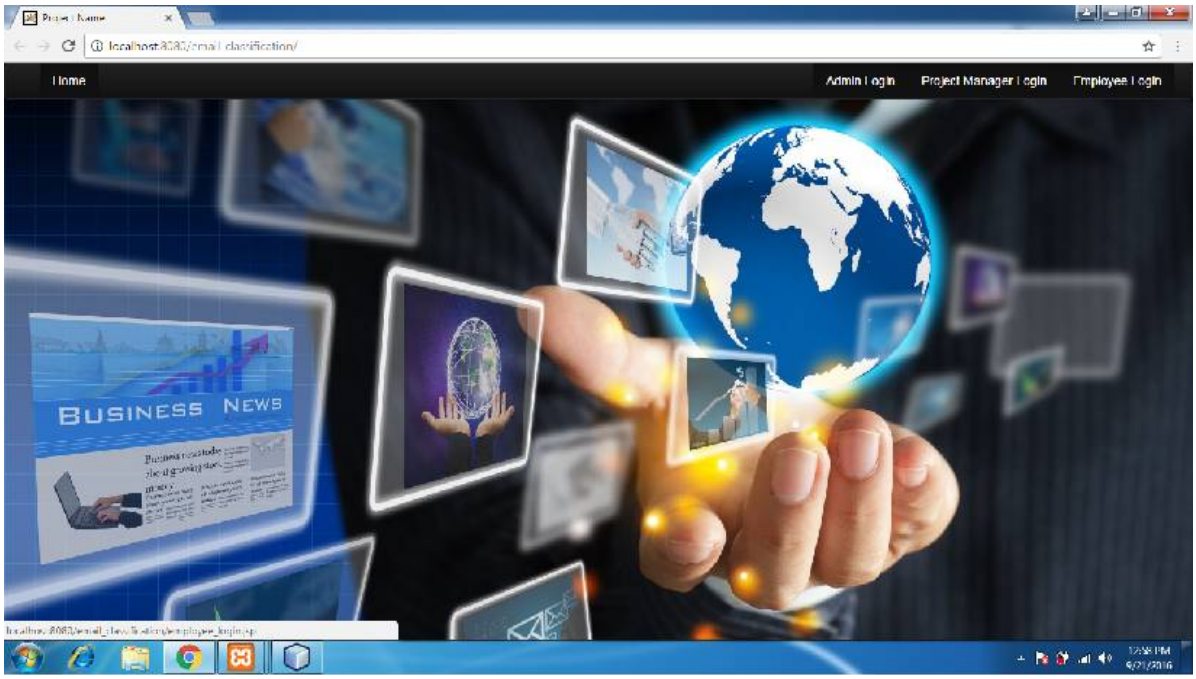


Fig.3. Shows front page

After this, Employee can send the mail from their registered mail id.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 9, September 2016**

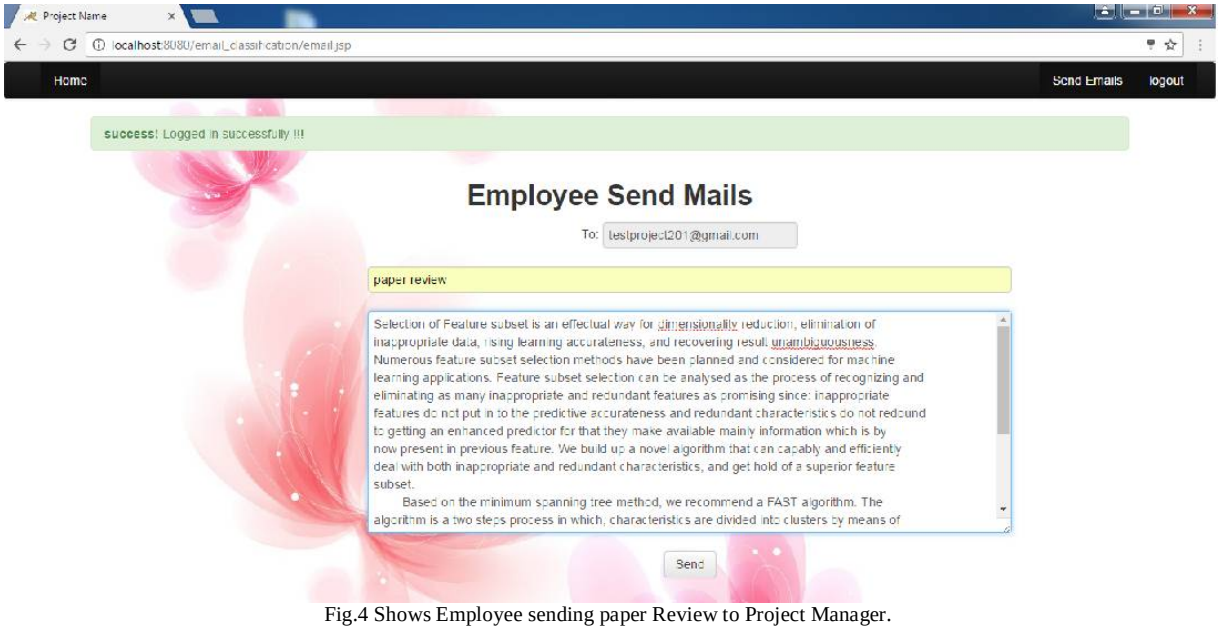


Fig.4 Shows Employee sending paper Review to Project Manager.

Project Manager then Retrives mail from his registered mail id which shows different category of mails in his or her inbox.

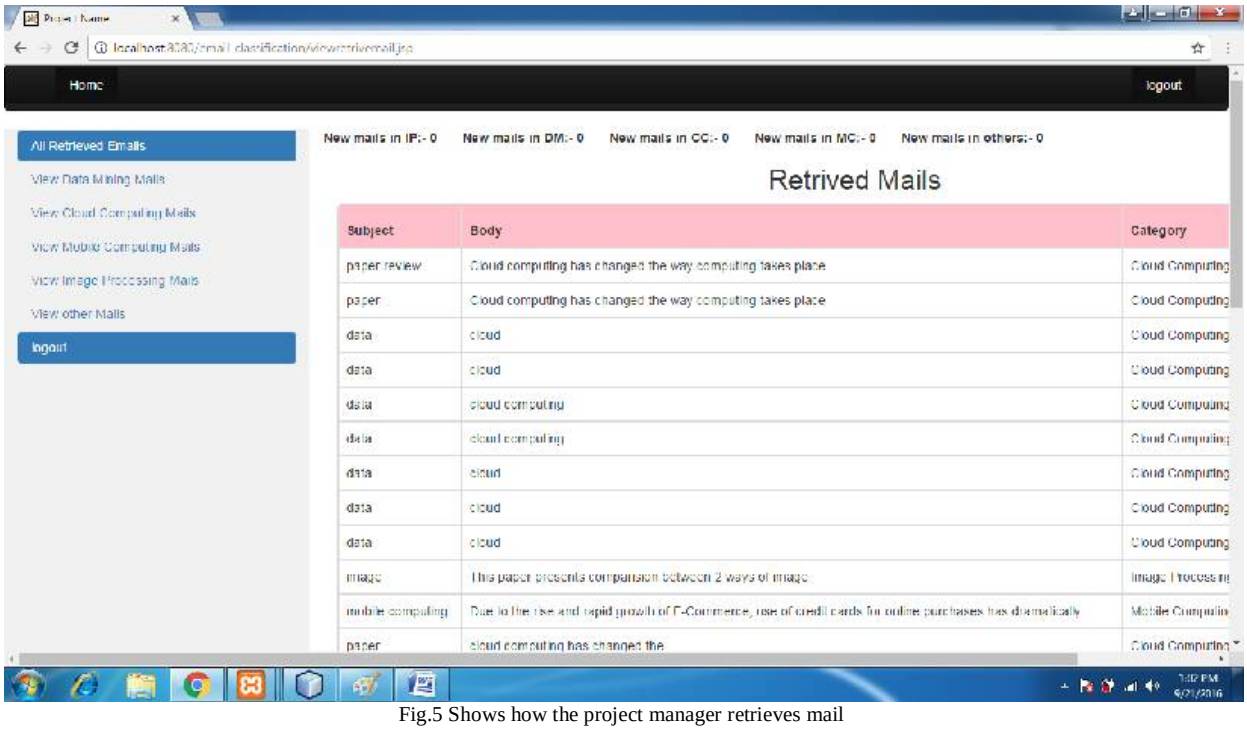


Fig.5 Shows how the project manager retrieves mail

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 9, September 2016**

Another part is of training in which, admin train the mail on basis of calculating probability of each word and then Mail can be categories into particular category like Data mining, cloud computing ,image processing, mobile computing. If mail arrived other than these categories then it will go to other category.

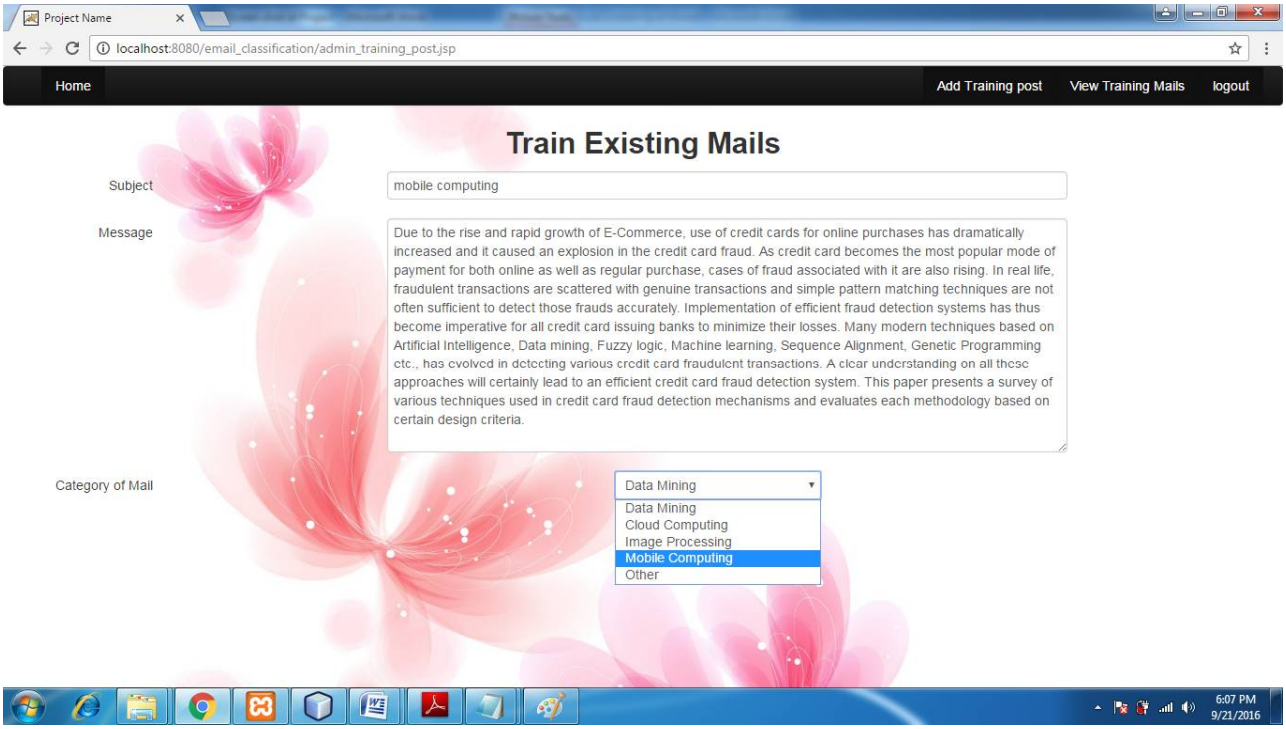


Fig.6 Shows Admin training the mails

## VI. CONCLUSION AND FUTURE WORK

This proposed application has discussed the importance of e-mail as a means of communication within engineering projects and even more so when a project becomes larger, increasingly multi- disciplinary and more distributed both spatially and temporally. Although high volumes of communication are seen to be indicative of a successful project and project progress, it is argued that there may be issues of information overload for engineers. In particular, the engineers classed as gatekeepers and information stars. E-mails also contain explicit rationale that can inform us 'why it is the way it is, yet the large volume of e-mail can present a challenge to aggregate and identify patterns Therefore, it has been proposed that being able to identify the purpose of the e-mail in real-time could aid engineers in their own Personal Information Management and aid in the identification of patterns/events within the engineering project leading to improvements in Project Management.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 9, September 2016**

## REFERENCES

1. Taehojo "Neural network for text categorization" school of information technology & Engineering, Ottawa University Canada Vol 2 issue April 2012
2. S.L. Ting,W.H.Ip Albert H.C. Tsang, "Is Naïve Bayes a Good classifier for document classification", International journal of software engineering and its applications Vol 5, No.3 July 2011
3. Guangxia Li,Steven C.H. Hoi,Kuiyu Chang,Wenting Liu,and Ramesh Jain "Collaborative online Multitask Learning" IEEE Transaction on Knowledge and Data Engineering.Vol 26 no.8,August 2014
4. StvanPilaszy,"Text Categorization and support vector machine",Department of measurement and information system Budapest university of technology and Economics
5. ThorstenJoschims "text Categorization with support vector Machines:learning many Relavent Features", university Dortmund, Germany
6. en.wikipedia.org/wiki/Naive_Bayes_classifier.
7. Harry zhang,"The Optimality of naive bayes", University of New Brunswick, China
8. James A. Gopsill , Stephen J. Payne & Ben J. Hicks, "An Exploratory Study into Automated Real-Time Categorisation of Engineering E-Mail", 2013 IEEE International Conference on Systems, Man, and Cybernetics.
9. Salwa Adriana Saab, Nicholas Mitri, Mariette Awad, "Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering", 17th IEEE Mediterranean Electrotechnical Conference, Beirut, Lebanon, 13-16 April 2014.
10. Aurangzeb Khan,Baharun B.Bahurdin,Khairullah Khan "An Overview of E-Documents Classification" International Conference on Machine Learning and Computing Vol.3 2011
11. C. Eckert, P. J. Clarkson, and M. Stacey, "Information flow in engineering companies: problems and their causes," *Design* Management: Process and Information Issues, Vol. 28, p. 43, 2001
12. J. Wasiak, B. Hicks, L. Newnes, A. Dong, and L. Burrow, "Understanding engineering email: the development of a taxonomy for identifying and classifying engineering work," Research in Engineering Design, Vol. 21, no. 1, pp. 43–64, 2010.
13. J. Wasiak, B. Hicks, L. Newnes, C. Loftus, A. Dong, and L. Burrow, "Managing by E-Mail: What E-mail Can Do for Engineering Project Management," Engineering Management, IEEE Transactions on, Vol. 58, no. 3, pp. 445–456, Aug. 2011.

## BIOGRAPHY

**Monali Khushal Patil** received bachelor's degree from Mumbai University in 2011. Now a P.G. student in Saraswati Education Society's Group of Institution Faculty of Engineering College, Mumbai University. Research interest in Data Mining.