Advanced Mining Approach for Classification of Emails Based on the Category

Vinita Thopte  
Department of Computer Engineering  
Zeal College of Engineering And Research  
Pune, India  
E-mail: vinitathopte1@gmail.com

Aishwarya Shete  
Department of computer Engineering  
Zeal College of Engineering And Research  
Pune, India  
E-mail: aishwryashete22@gmail.com

Pooja Todkar  
Department of Computer Engineering  
Zeal College of Engineering And Research  
Pune, India  
E-mail: poojasparkle124@gmail.com

Shweta Shinde  
Department of computer Engineering  
Zeal College of Engineering And Research  
Pune, India  
E-mail: shweta.shinde512@gmail.com

Prof. S. M. Kolekar  
Department of Computer Engineering  
Zeal College pf Engineering And Research  
Pune, India  
E-mail: sachinkolekar@zealeducation.com

Abstract: In daily life, Emails are the method or media between the computer users for communication purpose. Emails are used for both Professional and Personal levels. Every day the bulk of emails are arrived into user’s mail box. If user wants to search particular email from mail box the users have to search throughout the mailbox. So we wish to focus on the problem of categorizing mails into relevant category which would help to users to find important mail very quickly. In our paper, we will classify and customize our mail box. There are several options are present in our mail box like primary, social, promotions. Emails will save into relevant category. We will create some datasets like Shopping, Education, and Job etc. So whenever any mail will receive, it will check with the datasets and saved into corresponding category.

Keywords: Email classification, Email Customization, Naive Bayes Machine Learning, Clustering, Email Filtering

I. INTRODUCTION

The increasing the volume of the bulk emails. Email has become powerful medium for exchanging the information and communication for computer users.

It can reduce the complexity between the users. Emails are widely used for communication purpose between users.

Every day the bulk of mails are arrived into mail box so due to this if users want to search particular mail from his or her mail box then he have to go through whole mail box. So this is very time consuming process. In our project we will create datasets like Shopping, Education and Job etc. When email is arrived into mail box then it will check or compare with datasets then according to it will save into corresponding category.

II. RELATED WORK

A. A comparative study for some Content-based Classification Algorithms for Email Filtering

[1], The spam is major problem on the internet. In this paper, there is comparative study of Naive Bayes algorithm, Support Vector Machine, Local Machine Support Vector Machine. The motive of this paper was to reduce the computational cost and time.

B. Web spam classification techniques

[2], In this paper various classification algorithms are combined and used .The main aim is to take advantages of multiple classification schemes. In this paper Link based spamming and content based spamming is used.

C. Symbiotic filtering for spam email detection

[3], the paper is about symbiotic filtering for spam email detection. The main aim is to increase the user’s privacy and make the system robust against dictionary and contamination attacks .Here Naïve bayes algorithm and symbiotic filtering algorithm is used.

D. A review of machine learning approaches to Spam filtering [4].In this paper it is stated that the system is developed to filter the spam emails depending not only upon text emails but also upon the image contents in the email. For classification of emails having images as their contents logistic regression is used. For text classification Naive Bayes algorithm is used.

III. METHODOLOGY
Let S be the system defined as:
S= \{I/P, O/P, SS, SU, FS, A, DD, NDD\}

S= System.
I/P= Username, Password.
O/P= Classified Emails and Systematic mail box.
SS= System State.
    {Login State, Classification State, Result State}
DD=Deterministic Data.
    {Email will get classified successfully.}
NDD=Non Deterministic Data.
    {Emails with exceptional keywords.}
A=Set of Functions. = \{F1, F2, F3\}
F1= Preprocessing.
F2= Compare.
F3= Classify.
SU= Successful State.
    {Emails will be classified into relevant category.}
FS= Failure State.
    {Mail will be not classified or it will save into spam folder.}

In Our paper, we are using Naive Bayes algorithm For Classification of emails into relevant category. Naive Bayes Algorithm also called as text categorization and text retrieval. Normally this algorithm is used for supervised data.

Standard formula of Naïve Bayes for classification of the data is:

\[
P(C|X) = \frac{[P(X|C)P(C)]}{P(X)}
\]

Where,
P(C|X) = Posterior Probability
P(X|C) = Probability of predictor given keyword
P(C) = Prior probability of keyword
P(X) = Prior probability of predictor

Posterior Probability is the conditional probability which is calculated after the relevant evidences of the event.

Prior Probability is the unconditional probability which is assumes before any relevant evidences about event.

V. EMAIL CLASSIFICATION ARCHITECTURE

Email Filter Architecture is as follows:

When mail is arrived into mail box the tokenization will be done. It will create the tokens of the message. Then it remove the stop words like to, from, the etc. then keywords will compared with datasets and finding the probability of the datasets using naive bayes algorithm it will find the probability of each keyword. Highest the posterior probability that will be the results. The naive bayes algorithm is the best for finding the probability. According to that it will classified the mails into corresponding category. And user will get systematic mail box.

NAIVE BAYES:
Naïve bayes algorithm is used for calculating the probability. It also known as the text retrieval and text categorization. It can predict the probability based on the assumptions. It will find the posterior probability and according to that it will classified the text or information. Highest the posterior probability that will be the result.

**DATA FLOW DIAGRAM LEVEL 0.**

![Level 0 DFD](image)

Fig 2: Level 0 DFD

Level 0 DFD User will login with account. When the email is arrived it will compare with the datasets and it will be classified according to relevant category.

**DATA FLOW DIAGRAM LEVEL 1**

![Level 1 DFD](image)

Fig 3. Level 1 DFD

Level 1 DFD shows the flow of user mails is retrieved by the system to scan and filter the emails. After the authenticated user is logged in, all the incoming mails are retrieved by the system and the system classifies the mails into relevant category by checking the mail with datasets.

**VI. CONCLUSION**

In this paper, We consider the requirement of improving the efficiency of filtering techniques based on Naïve Bayes which is good Machine Learning algorithm for classification of the emails into relevant category. The project we concentrate on content of the emails. The result shows that our approach to classify the emails is reasonable and effective one. We will apply the classifiers for the email foldering task: Naïve Bayes. This approach of classification of emails will be feasible and easy for users to find the particular email form email box.

**REFERENCE**

It gives us great pleasure in presenting preliminary paper on "advanced mining approach for classification of email based on the category".

We would like to take this opportunity to thank our internal guide prof. s. m. kolekar for giving us all the help and guidance we needed. We are really grateful to her for her kind support. Her valuable suggestions were very helpful. We are also grateful to prof. sunil sangve, head of computer engineering department, Zeal College of engineering and research for his indispensable support, suggestions. in the end our special thanks to Prof. Nilesed paradeshi for providing various resources such as laboratory with all needed software platforms, continuous internet connection, for our project.

**REFERENCES**

[1]. Salwa Adriana Saab, Nicholas Mitri, Mariette Awad, „Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering”", Faculty of Electrical and Computer Engineering American University of Beirut, Lebanon, 2015


environment”. School of Computer Science and Engineering, South China University of Technology, China, 2014

[10]. Novino Nirmal A., Kyung-Ah Sohn, Tae-Sun Chung, “A Graph Model Based Author Attribution Technique for Single-Class E-mail Classification”, Department of Computer Engineering, Ajou University South Korea, 2015

[11]. Rizk Y., Mitri N., Awad M., “A Local Mixture Based SVM for an Efficient Supervised Binary Classification”, International Joint Conference on Neural Networks (IJCNN), Dallas, TX, August 4–9, 2013.

