An Automation Technique For Email Classification

Gaurav Bathe[1],Rohan Bapat[2],Nayan Tilekar[3],Akshada Shinde[4]

[1] *Gaurav Bathe, Computer Department, RMD SSOE, Maharashtra, India*
[2] *Rohan Bapat, Computer Department, RMD SSOE, Maharashtra, India*
[3] *Nayan Tilekar, Computer Department, RMD SSOE, Maharashtra, India*
[4] *Akshada Shinde, Computer Department, RMD SSOE, Maharashtra, India*

ABSTRACT

The email categorization has been proposed using Naive Bayes classification algorithm. .The categorization is based on not only the body but also the header of an email mes- sage. The metadata provide additional information that can be exploited and improve the categorization capability. Results of experiments on real email data demonstrate the feasibility of our approach.Results of system on real email data categorized into three types i.e. Primary, social and shopping. In particular, categorization based only on the header information is comparable or superior to that based on all the information in a message. The email communication becomes prevalent, all kinds of emails are generated.To classify emails for better visual representation and easy access to high priority important mails. The internal communications department of a company distributes an email message to all employees to remind the deadline of timecard submission.

Keyword : - *Bag of words,Naive trival(IR),Mail categorization*

1. INTRODUCTION

The emails have become an indispensable medium for people to communicate with each other nowadays. People can send emails not only to the desktop PCs or corporate machines but also to the mobile devices, and they receive messages regardless of the time and place. This has caused a drastic increase of email correspondence and made people spend significant amount of time in reading their messages. The email communication becomes prevalent, all kinds of emails are generated. People have got a tendency to make emails as their first choice when they need to talk to some- one.An email can be simply categorized into spams and non- spams.E-mails became the most important medium between individuals but also companies and various organizations and they settled down closely in almost any aspect of our everyday activity. The E-mails are not just simple text information; they can also transport different kind of attachments. E-mails are used in almost all areas of our life starting from regular activity at work, through shopping, advertising, logging in to various websites and services and ending with a private correspon- dence, just to give a few examples. Proposed system helps for categorization of the mails into the different categories. For example, we receives number of mails from social websites, from shopping sites ,from educational sites also some personalmails. That time this system helps to categorize the mails in different folders and also catogarized the subfolder i.e. Primary mails, Social mails and shopping mails.

2. LITERATURE SURVEY

An overview of email classification based only on not only the body but also the header of email messege.The classification of the email is a hierarchical system of categories used to organized the messege ,content in this

classifier.Email classification divide the different catogaries or subcatoga-ries.The classification of text catogarization is learning from high dimensional data.

2.1 Overview

”Toward Optimal Feature Selection in Naive Bayes for Text Categorization” This Paper present a novel and efficient feature selection framework based on the Information Theory, which aims to rank the features with their discriminative capacity for classification.

”Email Classification Research Trends: Review and Open Issues” This paper Personal and business users prefer to use email as one of the crucial sources of communication. The usage and importance of emails continuously grow despite the prevalence of alternative means, such as electronic messages, mobile applications, and social networks.

”Semi-Supervised Text Classification With Universum Learning” This paper devises a semisupervised learning withUniversum algorithm based on boosting technique, and fo- cuses on situations where only a few labeled examples are available

”A Bayesian Classifiers based Combination Model for Automatic Text Classification” Text classification deals with allocating a text document to a predetermined class. Generally, this involves learning about a class from representations of documents belonging to that class.

”M. T. Banday and S. A. Sheikh, Multilingual e-mail classification using Bayesian filtering and language translation, in 2014 International Conference on Contemporary Computing and Informatics” Personal and business users prefer to use email as one of the crucial sources of communication. The usage and importance of emails continuously grow despite the prevalence of alternative means, such as electronic messages, mobile applications, and social networks.

”W. W. Cohen, Learning rules that classify e-mail, in AAAI spring symposium on machine learning in information access Each email message is represented as vectors of features, in which the message body and each individual message header are represented as separate features.

”J. D. Brutlag and C. Meek, Challenges of the email domain for text classification, in ICML, It analysis the suitability with respect to email folder prediction and provide the baseline results.

”E. Blanzieri and A. Bryl, A survey of learning-based techniques of email spam filtering, Artificial Intelligence Review, vol. 29, pp. 63- 92, In this paper we give an overview of the state of the art of machine learning applications for spam filtering, and of the ways of evaluation and comparison of different filtering methods.

”T. S. Guzella and W. M. Caminhas, A review of machine learning approaches to spam filtering, Expert Systems with Applications, vol. 36, pp. 10206-10222, In this paper, we present a comprehensive review of recent developments in the application of machine learning algo- rithms to Spam filtering, focusing on both textual- and image- based approaches.

3. SYSTEM ARCHITECTURE

Results of system on real email data categorized into three types i.e. Primary, social and shopping. In particular, categorization based only on the header information is comparable or superior to that based on all the information in a message.

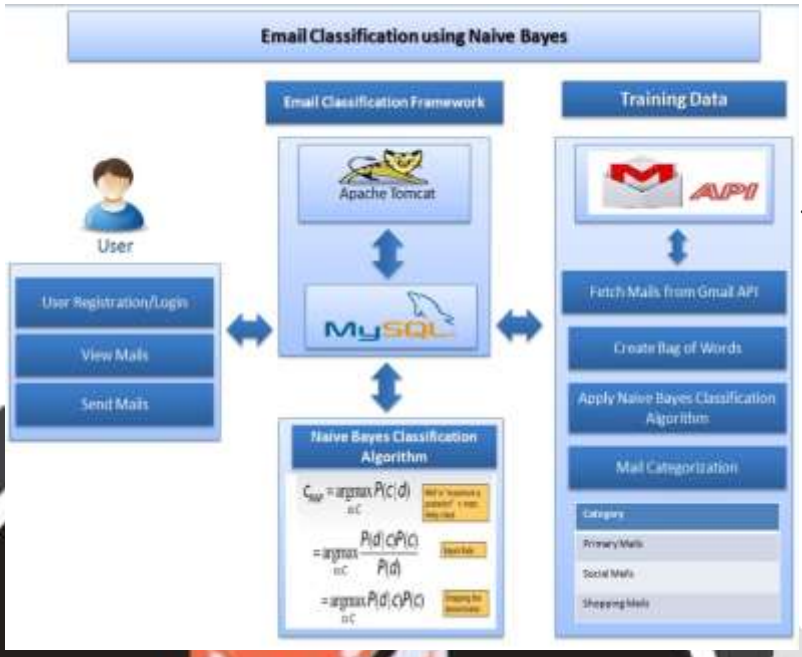

Fig -1: System Architecture

User Registration:
At the time of user registration user creates their id and password to our system.

Fetch Mails and Create Bag of Words:
Fetch mails from the Gmail API and store in database. The bag-of-words is a simplifying representation used in natural language processing and information retrieval (IR).

Gmail API: The Gmail API gives you flexible, Restful access to the user's inbox, with a natural interface to Threads, Messages, Labels, Fig1. Architecture of Email Classification using Nave Bayes Drafts, History, and. Settings. From the modern language of your choice, your app can use the API to add Gmail features.

Train Data:
For training the data apply Nave Bayes classification algorithm. Nave Bayes algorithm helps for classification of the mails in different categories.

Mail categorization:
The categorization is based on not only the body but also the header of an email message. The metadata (e.g. sender name, organization, etc.) provide additional information that can be exploited and improve the categorization capability. Primary Mails: It contains the personal mails.

Social Mails: It contains the social mails.
Shopping Mails: It contains the shopping related mails.
Education Mails:It contains the education related mails

4. ALGORITHM

The Bayes theorem says that the likelihood of prevalence might rely on the availability or non
Step 1: availability of another event. This dependency is written in terms of contingent probability. P (X—Y) = P (XY) / P (Y) P (Y/X) = P (YX) / P (X) P (XY) = P (X/Y) P (Y) = P (Y/X) P (X)

Step 2: The Bayesian classifier uses the theorem of Bayes theorem that says: P (Zj — d) = P (d—Zj) P (Zj) / P (d)

Step 3: Consider every attribute and sophistication label as a chance variable and given a record with attribute (X1, X2...Xn). The aim of this theorem is to predict category Z. we wish to seek out the worth of Z that maxi mizes P (Z—X1, X2...Xn).

Step 4:The approach taken is to reckon the posterior likelihood P (Z—X1, X2... Xn) For all worth of Z victimization the Bayes theorem. P(Z—X1,X2,...,Xn)= P(X1,X2,...,Xn—Z) P(Z) — P(X1,X2,...,Xn) therefore you selected the worth of Z that maximize P (Z—X1,X2...,An).this is similar to selecting the worth of Z that maximizes P(X1,X2,. ,Xn—Z)P(Z).

5. SUMMARY

In this paper, we proposed a system that categorize the emails using naive bayes algorithm.The main objective of the this to propose a solution to the problem of automatic classification incoming e-mails. System uses the Naive Bays Classification Algorithm for e-mail classification. Mails are mainly classifying into three category i.e. Primary mails, social mails and shopping mails.

6. CONCLUSION

The system proposes a solution to the problem of automatic classification of incoming e-mails. It categorizes the mails like social mails, primary mails and shopping mails with the help of Nave Bayes classification algorithm. Nave Bayes Classifier is amongst the most popular learning method grouped by similarities that works on the popular Bayes Theorem of Probability- to build machine learning models particularly for disease prediction and document classification.

7. ACKNOWLEDGEMENT

I am really glad to express my gratitude towards all who have contributed their views and opinions fro the comprehensive study of the research. I am really grateful to my guide Prof. Parth Sagar for his excellent guidance and support in the work. His suggestions and honest feedback which has helped me through out my research. I would also like to express my appreciation and gratefulness to Ms. Vina M. Lomte, Head of Department, Computer Engineering and to the Project Guide Mr.Shrikant Nagure I would lastly like to express my gratefulness and appreciation to all my colleagues who knowingly or unknowingly have encouraged me throughout my research.

8. REFERENCES

[1].”Toward Optimal Feature Selection in Naive Bayes for Text Categorization”
[2] ”Email Classification Research Trends: Review and OpenIssues”
[3] ”Semi-Supervised Text Classification With Universum Learning”
[4] ”M. T. Banday and S. A. Sheikh, Multilingual e-mail classification using Bayesian filtering and language translation,
[5] ”W. W. Cohen, Learning rules that classify e-mail,
[6] ”J. D. Brutlag and C. Meek, Challenges of the email domain for text classification,
[7] ”E. Blanzieri and A. Bryl, A survey of learning-based techniques of email spam filtering, Artificial Intelligence Review, vol. 29, pp. 63- 92,
[8] ”T. S. Guzella and W. M. Caminhas, A review of machine learning approaches to spam filtering, Expert Systems with Applications, vol. 36, pp. 10206-10222,
[9] ”S. Abu-Nimeh, D. Nappa, X.Wang, and S. Nair, A comparison of machine learning techniques for phishing detection,
[10] P. Pandian, K. Mohanavelu, K. Safeer, T. Kotresh, D. Shakunthala, P. Gopal, and V. Padaki, Smart vest: Wearable multi-parameter remote physiological monitoring system, Medical Engineering and Physics, vol. 30, no. 4, pp. 466477, 2008.
[11] ” Y. W. Wang, Y. N. Liu, L. Z. Feng, and X. D. Zhu, Novel feature selection method based on harmony search for email classification, Knowledge-Based Systems, vol. 73, pp. 311-323,