ISSN 2395-1621

E-mail Classification Using Text Mining

[#1]Sujata Borole, [#2]Sayli Arde, [#3]Anjali Jadon, [#4]Vaibhav Waghmare
[1]ssborole@gmail.com
[2]sayliarde@gmail.com
[3]anjalijadon889@gmail.com
[4]vaibhavw95@gmail.com

[#1234]Department of Computer Engineering MIT Academy of Engineering
Alandi, Pune -412105

ABSTRACT

Emails are nowadays used all around the world to communicate with each other. But there is some mails that are useful while some maybe just annoying and still you keep on receiving them. Such spam messages are causing serious trouble to Internet services as well as Internet users. This filtering required for this spam/ham is based on data classification. Hence we are proposing an email classifier which categorizes by identifying the main themes of an email by placing it into a pre-defined set of emails. Email Classifiers based on Bayesian theorem have been very effective in spam filtering due to their strong categorization ability and high precision. The purpose is to automatically classify mails into spam and legitimate message. The mails are classified on the bases of email body. We are therefore proposing a model for automatic sorting of spam and ham messages.

ARTICLE INFO

Article History

Received :3rd March 2016

Received in revised form :

4th March 2016

Accepted : 6th March 2016

Published online :

9th March 2016

I. INTRODUCTION

Electronic mail (or Email) - the method of exchanging digital messages from an author to one or more recipients that have contributed in better and faster way of communication. There are various categories in which the emails received by the user can be classified, like in Gmail there are forums, promotions, social, personal, spam, etc. But some mails can be unwanted as well as annoying. For example, after having a device-less and non-internet vacation at Maldives and checking your mail becomes so frustrating to know that you actually have received a large number of spam messages than the ham ones. Such mails have caused trouble a lot amongst the Internet users as well as for the Internet services.

Email spam are unsolicited bulk emails (UBE) or junk emails. Spam mails are used for spreading virus or malicious code, fraudulent banking, phishing, and advertising. So it can cause serious problem for internet users such as loading traffic on the network, wasting searching time of user and energy of the user, and wastage of network bandwidth etc. For fixing this problem we need filtering techniques based on data classification.

Filtering is nothing but arranging mails in specified order, such as spam and non-spam i.e. ham. Data Classification is a technique of data mining for identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Data mining is defined as a discovering useful knowledge from data. Various applications of data mining are sales transactions, stock trading records, product descriptions, sales promotions, company profile and performance, medical and health industry, and customer feedback, reporting online analytical processing, business performance management and so on. Classification is the process of finding model that describe and distinguishes data classes or concepts. Classification consists of two steps. First is process learning step: where a classification model is constructed and second classification step: In this step the model is used to predict class labels depending on the learning step for given data. The need for a spam filter is needed but only research on the particular topic gives us the right method to implement this model. The next section describes the Literature Survey of our topic.

II. LITERATURE SURVEY

Text mining being a vast subject is not that easy to apply. "Text Mining Process, Techniques and Tools: an Overview by Vidhya. K. A \& G. Aghila" helped us to understand the text mining process and techniques. Basically it explained us the concepts of information retrieval and extraction with effective knowledge discovery. As we wanted to search more relevant data about which technique and algorithm to use, we referred "Email classification for Spam Detection using Word
Stemming, 2010 International Journal of Computer Applications" paper with other papers such "Email Classification Using Data Reduction Method by Rafiqul Islam and Yang Xiang" and "MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION, International Journal of Computer Science & Information Technology (IJCSIT), Feb 2011". These papers helped us to search and study different email classification algorithms. Machine learning algorithms have achieved a higher success rate. In the later paper, we studied about such machine learning algorithms - Bayesian classification, k-NN, ANNs, SVMs, Artificial immune system and Rough sets. Also we learnt about their applicability to the spam classification. We have taken the idea of spam filtering by Naive Bayesian theorem from our base IEEE paper "An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques by Vikas P. Deshpande, Robert F. Erbacher, and Chris Harris". The paper by International Journal of Advanced Research in Computer Science & Software Engineering "Effective Email Classification for Spam and Non-Spam by Savita Pundalik Teli, Santoshkumar Biradar" gave us the detailed idea of how the model for spam filtering may work. It includes the classification, spam filtering process and the how the Bayesian theorem relates to the proposed model. Apart from all the above papers we have referred many other papers and websites as well. Following are the few algorithms that we studied for our literature survey.

A. k-nearest neighbour algorithm

The k-nearest neighbour problem arises in several applications such as density estimation, pattern classification and information retrieval. The problem is to find, among a set of points (or feature vectors), the one which is most similar or closest to a given test point according to some dissimilarity or distance measure.

Advantages: Robust to noisy training data. Effective if training data is large.

Disadvantages: Need to determine value of parameter k. Computation cost is quiet high.

B. Artificial Neural Networks (ANNs) algorithm

Artificial neural networks (ANNs) are a form of artificial intelligence which attempt to mimic the behaviour of the human brain and nervous system.

Advantages: Neural networks are quiet simple to implement. They often exhibits patterns similar to those exhibited by human.

Disadvantages: Neural networks cannot be retrained, it is impossible to add data to an existing network.

C. Decision Tree algorithm

Decision Tree works for classification and regression problems. So if to predict categorical data like (Red, Green, up, down) or if data is continuous like 2.9, 3.4 etc. Decision Tree is able to handle both kind of data (continuous as well as discrete).

Advantages: The coolest thing about Decision tree is only a table is required that will build a classifier directly from that data. And it doesn't consider the properties that are not required.

Disadvantage: The costs involved in such training makes decision tree analysis an expensive option. Among the major disadvantages of a decision tree analysis is its inherent limitations.

D. Naïve Bayes algorithm

Naive Bayes is a simple technique for constructing classifiers. Here value of particular feature is independent of the value of other feature, given the class variable. There is no correlation between the one feature and the other feature to predict class labels.

Advantages: Easy to implement. Requires small amount of training data to estimate the accuracy.

Disadvantages: As one feature is independent of other feature the independency accuracy is lost. No Dependency exist among variable.

III. PROPOSED SYSTEM

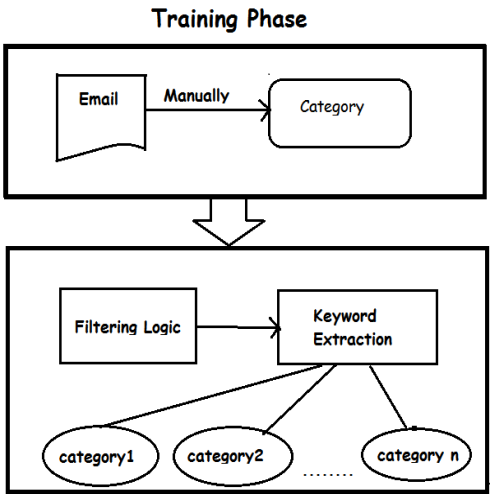
The email service being popular has its own disadvantages too. Emails that we receive are classified into different categories according to the context they contain. We are trying here to put forth this idea in our project.
Spam mails are causing a lot of trouble for the users. The idea is to classify the mails into spam and ham or non-spam mails. An email consists of two parts - message header and message body. The header consists of sender and receiver address, subject, date and server address, etc. While the message body contains the text data. Hence excluding the header, we will now concentrate on the body of the message. As we mentioned before, the message body consists of text, this text is divided into set of tokens. These tokens are separated by blank spaces or any English punctuation marks. The proverbs, stop words, html tags, etc. are all removed and the filtering of the tokens ends here. The model's process consists of training as well as classification process, where the keyword dataset is built and automatic sorting is done with the help of the Bayesian Classifier. Each unclassified goes through the filter and then through the classifier thus sorting the mail into the pre-defined categories. Bayesian filtering is named after the English Mathematician Thomas Bayes who developed the theory of probability inference. Bayesian filters being adaptable can train themselves to identify new patterns of spam and hence

can be adaptive to the user's specific parameters for identifying spam. These filters take the whole context of a message into consideration, for example, not every e-mail with the word "free" in it is spam, so the filter verifies the e-mail with the word "free" by calculating the probabilities. Obviously, the productiveness of spam filtering depends very much on the message database used for learning. So, to make it effective for a particular user it is necessary that the spam filter should learn from the message database of this user and automatically make changes when filtering is incorrect. Thus, possibility to customize spam filtering immediately after the installation and to correct filtering mistakes in order to avoid them in the future make the Bayesian Filter more advantageous. We are proposing to design this system in Java programming language as Java runs on a variety of platforms, such as Windows, Mac OS, and the various versions of UNIX. The compatibility of Eclipse or Netbeans IDE with Java serves us favourable results.

IV. DETAILED DESIGN

The whole email classification process is divided in two phases as given below:

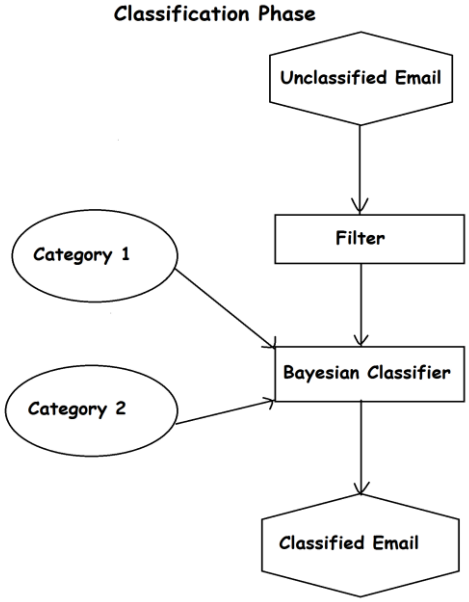
A. First phase - Training phase/Learning phase - To proceed with automatic sorting, we need a pre-defined keyword dataset. For this we need to manually sort the emails and extract the common keywords from the spam mails to the keyword dataset. The above procedure takes place in the Learning/Training Phase. During this phase, the classifier is trained to recognize the keywords for each category, so when a new mail arrives automatic keyword recognition and sorting will take place using the pre-defined keyword dataset. The message body consists of text, this text is divided into set of tokens. These tokens are separated by blank spaces or any English punctuation marks. The proverbs, stop words, html tags, etc. are all removed. Hence the training phase is responsible for keyword dataset building.



Algorithm:-

1) To specify each mail its category, manually.
2) To separate the header and body of the mail.
3) To divide the mail (message) body into tokens.
4) To filter out the stop words such as html tags, articles, proverbs, noise words.
5) And finally, to extract keywords and store them along with frequency count in the keywords database of the selected categories.

B. **Second phase - Classification phase -** Now that the learning is been done, we may proceed further. The new mails received now should be assigned to respective categories. Basically, here we compare the contents of the mail with those of all category keyword databases using Bayesian theorem and look for best matching category for the mail. The new mail undergoes the same process of breaking down into tokens and then filtering. Then these tokens are compared with the keyword database that we built in the previous phase. The probability that the mail belongs to a category is found out for all categories. The probability of the matching category is the compared with the same category's threshold and is truly classified into spam/ham.



Algorithm:-

1) Now apply the task of separating the mail header and body and then dividing the body into tokens to the newly arrived mails.
2) Filter the stop words and extract the keywords.
3) Find the probabilities for all the categories.
4) Find the category with the highest value of the obtained probabilities.

V. CONCLUSIONS

We have put forth the survey and working of the e-mail classification process using Bayesian Theorem. Considering the requirements for efficiency and accuracy along with the literature survey, there is for sure lots of future work to do. We may consider more wider approach for keyword dataset like emphasizing on the semantics and not just on individual

words, making extraction efficient, increasing the number of categories and so on.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of various authors who have researched and written their theories for this topic. We would also like to acknowledge Prof. Kavitha S. for her guidance and suggestions.

REFERENCES

[1] An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques BY Vikas P. Deshpande, Robert F. Erbacher, and Chris Harris, Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY 20-22 June 2007.

[2] An Approach to Email Classification Using Bayesian Theorem By Denil Vira, Pradeep Raja & Shidharth Gada, Global Journal of Computer Science and Technology Software & Data Engineering.

[3] Effective Email Classification for Spam and Non-Spam by Savita Pundalik Teli, Santoshkumar Biradar, International Journal of Advanced Research in Computer Science and Software Engineering, June 2014,

[4] Effective Pattern Discovery In Text Mining,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.

[5] MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION by W.A. Awad and S.M. EL seuofi, IJCSIT, Feb 2011.

[6] Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm, IJITEE.

[7] A Survey of Text Mining Techniques and Applications, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE.

[8] Survey Paper on Document Classification and Classifiers, IJCST.

[9] Text Mining Process, Techniques and Tools: An Overview, International Journal of Information Technology and Knowledge Management.

[10] Email classification for Spam Detection using Word Stemming, 2010 International Journal of Computer Applications.

[11] Email Classification Using Data Reduction Method by Rafiqul Islam and Yang Xiang.

