A Hybrid E-Mail Spam Filtering Technique using Data Mining Approach

Subhana Khan
Department of Computer Science Engineering
Medi-Caps Institute of Technology & Management,
Indore, M.P, India

Pramod S. Nair
Department of Computer Science Engineering
Medi-Caps Institute of Technology & Management,
Indore, M.P, India

Abstract - Communication is a primary need of human beings therefore new techniques are invented to support low cost, efficient and adoptable techniques for new generation communication technology. SMS and email messages are such techniques which are adoptable, efficient and cost effective. Now-a-days most of the communications, official and personal are performed using email messaging. But these mails are not much secure due to untrusted networks and intermediate network attackers. Therefore a number of techniques are developed for securing the email messaging from the attackers among them the spam filters are an essential contribution. In the proposed work the traditionally available techniques of spam filtering are investigated. In addition to that a new technique using hybrid methodology is presented. The proposed technique incorporates the Bayesian classifier and the neural network. In the result section the performance of the proposed technique is compared to the traditional Bayesian classifier.

Keywords – Email Classification, Text Mining, Hybrid Classifier, Bayesian Classifier, Neural Network.

I. INTRODUCTION

In 21st century email become one of the most important parts of our life whether it is personal or professional. A professional person checks its email daily related to its work. Emails are the cheap mean of communication. It is very efficient and consumes less time, it means it is rapid in nature. All of these reason it become popular in a short period of time. Email spam, we heard this name many times related to emails. Spam emails are the unwanted emails which are usually send for their interest. They may be of any nature, whether it is an advertisement or a simple text email. Some spam emails contains only link, for another web page, in this navigation they can steal our personal details like credit card number, bank account number and so on. Generally these emails are sending in a bulk. Spamming of emails grown rapidly since early 1990. A study shows that 80% of spam emails are send through a network of virus infected computers. The legal status of spam is changing from one jurisdiction to another therefore no strong rule is adopted against spammers [1].

We always think that how spammers collect large databases of recipients. They visit chat rooms, websites, customer lists and newsgroups, so from all these places they collect email id of the recipients. Viruses also play a key role in this. They scan address books of users and sold it to the spammers these address books are helpful in getting email addresses. They also use a trick to know email addresses of users from their known information like postal address and this trick is known as email epending. In the first half of 2010, around 88-92% of spam emails were sent according to the Message Anti Working Group.

II. LITERATURE SURVEY

Kleanthi Georgala et.al. introduces a method that deals with unwanted mail messages by combining active learning with incremental clustering [6]. The proposed approach is motivated by the fact that the user cannot provide the correct category for all received messages. The email messages are divided into chronological batches (e.g. one per day). The user is asked to give the correct categories (labels) for the messages of the first batch and then proposed algorithm decides when to ask for a new label, based on a clustering of the messages that is incrementally updated. Author test different variants of the algorithm on a number of different data sets and show that it achieves very good results with only 2% of all email messages labeled by the user.

Yanyan Guo et.al. introduces Bayesian spam filtering is a classification method based on the theory of probability and statistics, and the Bayesian spam filtering based on MapReduce can solve the defect of the traditional Bayesian spam filtering that consumes large amounts of system resources and network resources when the mail set is pre-training [7]. It needs to classify mails manually in the pre-training phase of mail set, which consumes a lot of human and financial resources and affects the efficiency of the system. Bayesian spam filtering mechanism based on the decision tree of the attribute sets dependence in the MapReduce framework which is presented in this paper. And the decision tree of attribute sets dependence is used in the training stage of the mail set, which improves execution efficiency of the system by lowering the time complexity.

Jun Liu et.al. analyses several common algorithms for spam filtering and shows the advantages and disadvantages of these algorithms for spam filtering [11]. Each algorithm is only suitable for filtering specific spam. Some algorithms are suitable for Chinese, and some algorithms perform well in English. In a lot of spam, it is not reliable and inefficiency to using a single algorithm to separate out spam. Moreover, an intelligent method that it has the ability of self-learning by using the contents of the e-mails is introduced. Finally, the outcome of experiment shows that the intelligent method achieves a better efficiency and performance.

III. PROPOSED WORK

The given section provides the detailed description of the proposed methodology by which the email data is classified effectively and more accurately. Therefore the two different classifiers are utilized to get a hybrid classification algorithm. The key concept of the proposed algorithm design is depends on the need which accept the data D as input and produces the classification label in two different labels. Thus the proposed classifier working as a function in the following manner:

$$f(D) \rightarrow \{spam, legitimate\}$$

The proposed data model for learning and classification of spam and legitimate emails are described in the given figure 1. There components and subcomponents with their working are explained in the following manner:

Input Learning Set

The proposed data model is a supervised hybrid data modeling and learning technique, therefore the learning required using a set of samples of legitimate samples and spam samples. Therefore a significant amount of files with their Meta data information is provided as input to the system. The Meta data contains the file name and their class labels for making successfully training.

Data Pre-Processing

The input training samples are pre-processed. In pre-processing the data is first clean for processing. Therefore the input files are processed with the pre-defined stop words removal process, in this process the frequent words are removed and similarly the punctuations from the input samples are also removed.

Tokenization

After the pre-processing of the input files the data is tokenized, during tokenization similar words in both kinds of samples are counted.

Bayesian Classifier

The Bayesian classifier is a probabilistic classifier and it is also works as a statistical function which accepts the tokenized data and their count in spam files and also for the legitimate files therefore the similar word probability is computed for both the kinds of input samples for generating the following:

- spam probability
- legitimate probability

In order to understand the spam and legitimate probability the following example can help:

Suppose a spam file contains a string such that:

Hello sir, you get a credit card from the bank

And for the legitimate file:

Hello sir, we are from the bank making effort to serve you better regards bank

Here the word bank is common in both the samples thus according to their occurrence in both the files are estimated and in spam file the word is found only one time and in legitimate email it is obtained two times thus according to their contribution in the email the probability distribution is counted and listed.

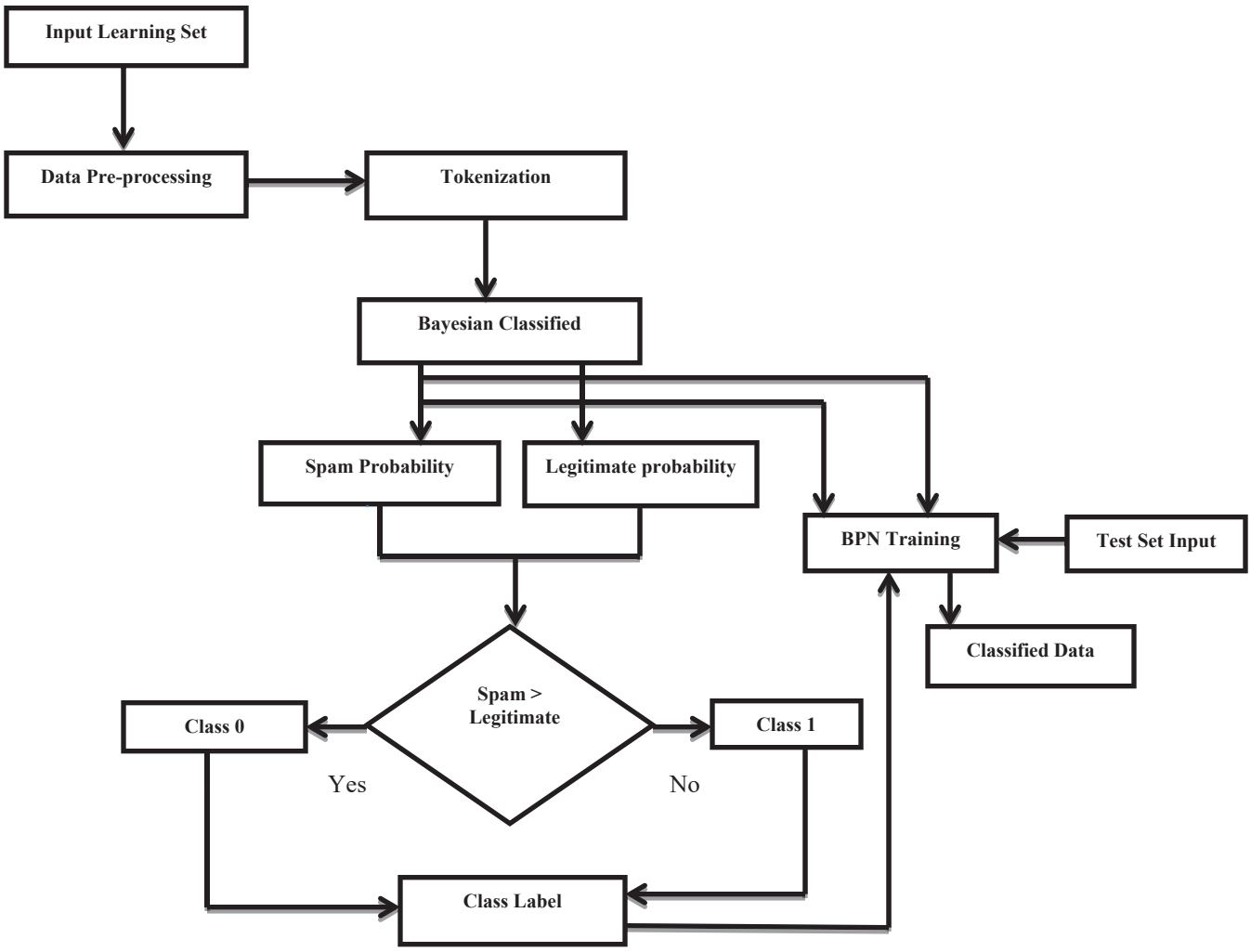

Figure 1. Proposed System Architecture

Class Labels

The class labels are the key element in supervised learning thus for deciding the class labels for the specified word in spam email and legitimate email the probability distribution for both the email words are compared and the probability in finding a spam email is higher than the legitimate email and after that the class is defined as the spam email.

Thus a combined input for learning to the BPN algorithm is prepared for each individual words is prepared and can be simulated using the table 1.

Table 1. Training Samples of BPN Algorithm

Word	Spam probability	Legitimate	Class label
Bank	0.125	0.253	1

BPN Training

The back propagation algorithm is executed and accepts the probability to be spam and legitimate for target values 0 or 1. In such conditions for each input samples the back propagation algorithm is trained and prepared for classification.

Test Set Input

In this phase the trained back propagation algorithm is utilized for identifying their class labels. Additionally for the entire input email contents the aggregate class label is computed that produces the final class labels for individual files.

Classified Data

It is an aggregate class labels that represent the contents of the input files are either spam email or legitimate email

IV. RESULT ANALYSIS

This section explains the performance outcome and their analysis thus a number of different parameters are listed and according to the evaluated performance the system outcomes are demonstrated.

- *Accuracy*

The amounts of data that are correctly identified during the classification of emails are known as the accuracy. Calculation of accuracy can be evaluated using the following formula:

$$\text{accuracy} = \frac{\text{total correctly classified data}}{\text{total input data to classify}} X100$$

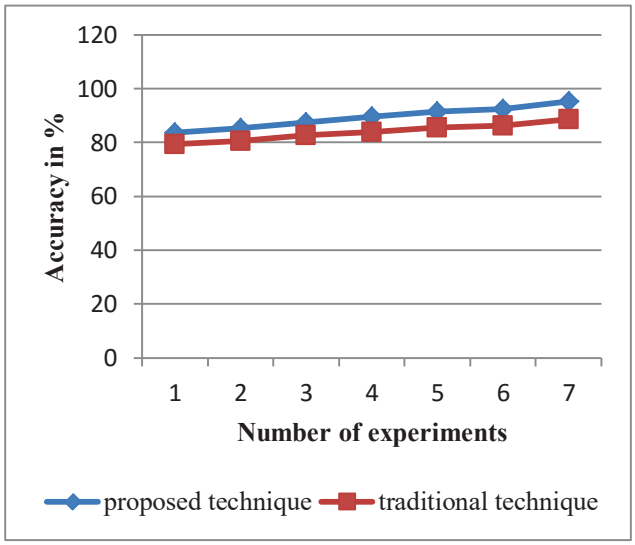

Figure 2. Accuracy of Spam Filter Algorithm

Figure 2 represents the percentage of accurate spam filter recommendation. Where X axis of graph shows the number of experiments and the Y axis shows accuracy in percentage. The red line shows the performance of the traditional approach and the proposed technique is demonstrated using blue line. According to the comparative

results the performance of the proposed algorithms is highly accurate and efficient as compared to the traditional algorithm.

- *Error Rate*

Error rate is the ratio of total incorrectly identified samples to the total input samples for the classification technique. Calculation of error rate can be evaluated using the following formula:

$$\text{error rate} = \frac{\text{total incorrectly identified samples}}{\text{total input samples}} X100$$

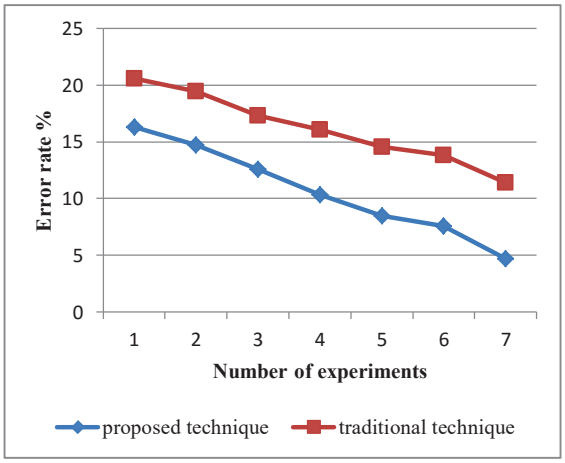

Figure 3. Error Rate of Spam Filter algorithm

Figure 3 represents the percentage of error rate of spam filter algorithm. Where X axis of graph shows the number of experiments and the Y axis shows error rate in percentage. The red line shows the performance of the traditional approach and the proposed technique is demonstrated using blue line. According to the comparative results the error rate of the proposed algorithms is less as compared to the traditional algorithm.

- *Time Consumption*

The amount of time required to process the data using the selected algorithm is known as time consumption of the system.

Calculation of time consumption (in sec) for spam filter algorithm:

The figure 4 shows the comparative time consumption of both the algorithms. According to the graph X axis shows the number of experiments performed with the system and the Y axis shows the amount of time consumed in seconds. According to the simulated results the performance of the proposed technique consumes more time as compared to the traditional classification technique. Thus in the parameter of time consumption or time complexity the performance of traditional classifier is more appropriate than the proposed technique.

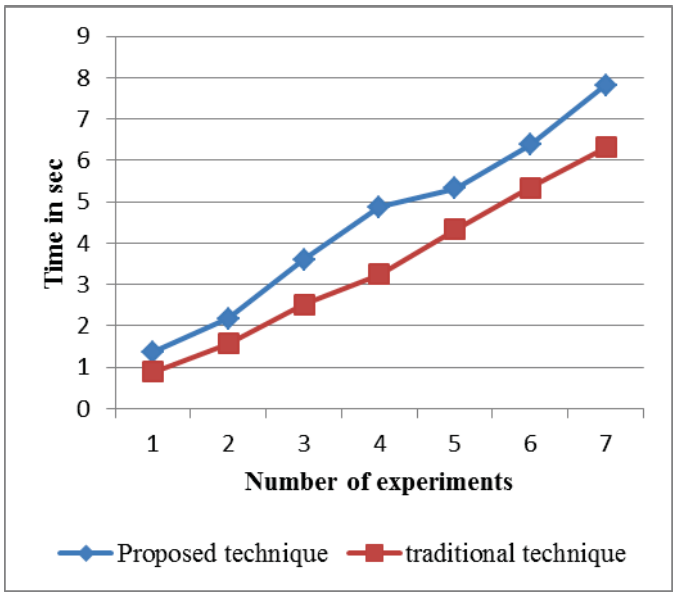

Figure 4. Time Consumption of Spam Filter Algorithm

- *Recall*

The search recall values are measured in this section, that is an accuracy measurement in terms of relevant document retrieved according to the input search query. This can be evaluated using the following formula.

$$recall = \frac{relevantdoucment \cap retrieveddocuments}{relevantdocuments}$$

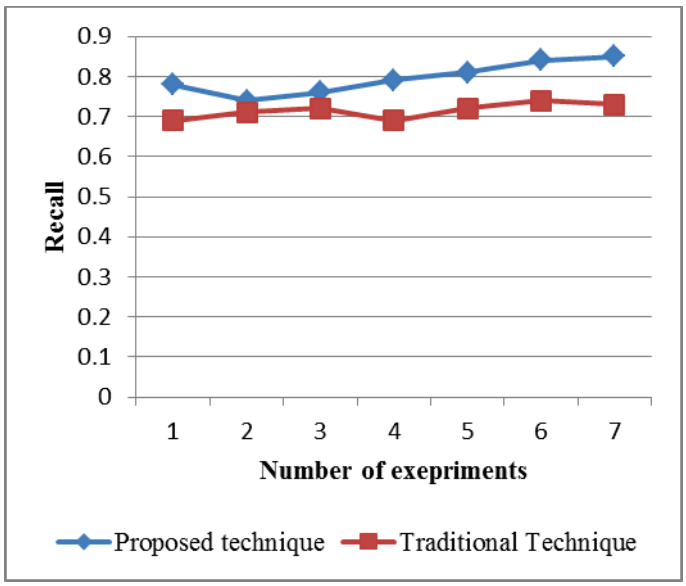

Figure 5. Recall Rate of Spam Filter Algorithm

Figure 5 represents the percentage of recall rate of the spam filter algorithm. Where X axis of graph shows the number of experiments and the Y axis shows the recall rate. The red line shows the performance of the traditional approach and the proposed technique is demonstrated using blue line. According to the comparative results the performance of the proposed algorithms is much adoptable as compared to the traditional algorithm.

V. CONCLUSION

In this presented work the spam email filtering techniques are investigated and some additional issues are addressed. The key issue of the spam filtering is to classify similar contents into more than two classes. Thus in this presented scheme a new model for classify the spam mails are presented. The proposed data model includes the hybrid approach of classification therefore two classifiers namely Bayesian classifier and back propagation neural network is organized together to enhance the accuracy of traditional classifiers.

The implementation of the proposed classification technique is performed using the JAVA technology and their performance is computed in terms of time consumption, error rate, accuracy and recall. The performance of the proposed classification technique found more optimum as compared to the traditional approach of classification. The performance of the system is reported using a summary table as given in table 2.

Table 2. Performance Summary of Spam Filter Algorithm

S. No.	Parameters	Proposed Technique	Traditional Technique
1	Time consumption	High	Low
2	Accuracy	High	Low
3	Error rate	Low	High
4	Recall	High	Low

According to the obtained results in table 2, the performance of the proposed classification technique is found optimum and efficient as compared to the traditional technique. Thus proposed model is more adoptable as compared to the traditional approach of classification.

VI. FUTURE WORK

The proposed work is intended to develop an accurate and efficient approach of classification and using hybridization of two different algorithms Bayesian classifier and back propagation neural network that is achieved. But the performance of the system in terms of time consumption is lacked thus in near future the work is extended for improving the time complexity of the proposed data model.

REFERENCES

[1] S. Atskins, “Size and cost of the problem”. In Proceedings of the Fifty-sixth Internet Engineering Task Force (IETF) Meeting, (San Francisco, CA), SpamCon Foundation, March 16-21-2003.

[2] Ann Nosseir, Khaled Nagati and Islam Taj-Eddin, “Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.

[3] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, “Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection”, DELIS – Dynamically Evolving, Large-Scale Information Systems, 2006.

[4] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Second Edition.

[5] Jerzy W. Grzymala-Busse and Ming Hu, “A Comparison of Several Approaches to Missing Attribute Values in Data Mining”,Springer-Verlag Berlin Heidelberg, pp. 378−385, 2001.

[6] Kleanthi Georgala, Aris Kosmopoulos, George Paliouras, “Spam Filtering: an Active Learning Approach using Incremental Clustering”, WIMS Thessaloniki, Greece Copyright is held by the owner/author(s).Publication rights licensed to ACM, June 02 - 04 2014.

[7] Yanyan Guo, Lei Zhou, Kemeng He, Yuwan Gu and Yuqiang Sun, “Bayesian Spam Filtering Mechanism Based on Decision Tree of Attribute Set Dependence in the MapReduce Framework”, The Open Cybernetics & Systemics Journal, Vol. 8, pp. 435-441, 2014.

[8] Zhiqiang Ma, Rui Yan, Donghong Yuan and Limin Liu, “An Imbalanced Spam Mail Filtering Method”, International Journal of Multimedia and Ubiquitous Engineering Vol. 10, No. 3, pp. 119-126, 2015.

[9] Sin-Eon Kim, Jung-Tae Jo, Sang-Hyun Choi, "A Spam Message Filtering Method: Focus on Run Time", Advanced Science and Technology LettersVol.76, pp.29-33, CA 2014.
[10] Tarek M Mahmoud, Alaa Ismail El Nashar, Tarek Abd-El-Hafeez and Marwa Khairy, "An Efficient Three-phase Email Spam Filtering Technique", British Journal of Mathematics & Computer Science Vol. 9, No. 4, pp. 1184-1201, 2014
[11] Jun Liu, Shuyu Chen, Kai Liu and Yong Zhou, "A Composite Intelligent Method for Spam Filtering", International Journal of Security and its ApplicationsVol.8, No.4, pp.67-76, 2014.
[12] Ya-ping Jiang, Yue-xia Tian and Xiao Mei, "A spam filtering model based on immune mechanism", Journal of Chemical and Pharmaceutical Research, Vol. 7, No. 6, pp. 2533-2540, 2014.
[13] Kamini (Simi) Bajaj and Josef Pieprzyk, "A Case Study of User-Level Spam Filtering", CRPIT Volume 149 - Information Security, 2014.
[14] Fatiha Barigou, Bouziane Beldjilali, and Baghdad Atmani, "Using Cellular Automata for Improving KNN Based Spam Filtering", The International Arab Journal of Information Technology, Vol. 11, No. 4, July 2014.
[15] Noormadinah Allias, Megat Norulazmi Megat, Mohamed Noor, Mohd. Nazri Ismail, "A Hybrid Gini PSO-SVM Feature Selection Based on Taguchi Method: An Evaluation on Email Filtering", IMCOM (ICUIMC), Siem Reap, Cambodia, January 9–11, 2014.