This full-text paper was peer-reviewed and accepted to be presented at the IEEE ICCSP 2015 conference.

Content Based Spam Detection in Email using Bayesian Classifier

Sunil B. Rathod, Tareek M. Pattewar

Abstract— **Internet provides Emails as means of data communication. Email messaging is an essential contribution. Hacking attacks, phishing attacks and malicious attack are frequently undergo email services to attempt fraud and deception motivation. They use emails to obtain personal credentials of user for financial gain. Emails with genuine content may include phishing URLs for stealing of useful data such kind of emails are nothing but a spam.**

In order to detect and filter such kind of emails. Bayesian classifier is used and The performance of Bayesian classifier is evaluated in terms of Accuracy, Error, Time, Precision and Recall, Bayesian classifier is used for email classification and detection of spam mails.

Index Terms— Bayesian Classifier, Email Classification, Spam Filtering.

I. INTRODUCTION

Email services are becoming popular by means of information communication over Internet. These Email facility has also created troubles to user through Electronic junk mails. These are called as Spam mails .The spam mails are sent to many users in bulk as advertising mails, claim mails . Some of the mails contains genuine content with malicious URLs called as phishing mails. It also spread virus, malicious attacks through spam mails. Spam mails are also said to be Unsolicited Bulk Emails and Its another part is Unsolicited Commercial Emails. Hacker, phishers and malicious attackers are frequently using email services to send false kinds of messages by which target user can loss their money and social reputations. These results into gaining personal credentials such as credit card number, passwords and some confidential data.

To stop such kind of things one should employ following : Do not reply to spam mails, Do not click the links / URLs in Emails, Do not post your email ids on the unrecognized web sites.

Content Based Spam Filter :

Content based filter checks for Text in the body of Email, then URL and It also considers the mail header like subject for classification of text . It performs Text classification task by employing preprocessing on TEXT in terms of HTML tag removal, Stopword Removal, Tokenizing and Word frequency calculation for determining word probability to find out whether a given mail is spam or not.

The rest of the paper is organized as follows next section describes the related work concerning spam classification and filtering method. The section III discusses on algorithm study, We have used Bayesian classifier algorithm and evaluation criteria for our work. The section IV elaborates experiment which consists of training dataset, preprocessing, application of Bayesian classifier and testing dataset then performance evaluation and The section V discusses about result which we have derived by considering performance measurement parameters. Finally the section VI concludes the paper.

II. RELATED WORK

The existing work undergo an implementation on detection of malicious URL in Email by Dhanalakshmi R and Chellapan C 2013, they considered Age of domain, Host based features, Lexical features and Page rank for analysis of URL to classify into malicious URL and legitimate URL. They have used Bayesian classifier to improve the accuracy by reduced feature sets and considered phishtank dataset, The work was restricted to URL in Email only [1].

Sahami et al. 1998, has given a spam classification method using a Bayesian approach. A Bayesian classifier is statistical classifier works on independence computation of probability. They have considered content of e-mail with features of domain, and shown that accuracy can be increased [2].

V Christina et al, had shown that the need of effective spam filters increases. He discussed spam and spam filtering methods and their correlated problems [3].

Sadeghian A. et al, had presented spam detection based on interval type-2 fuzzy sets. This system gives user more control on categories of spam and permits the personalization of the spam filter [4].

Congfu Xu et al. has derived a feature extraction on Base64 encoding of image with n-gram technique. Effectiveness and efficiency in detecting spam images are shown by these features from legitimate images by training a SVM.

Sunil B. Rathod is the PG Student, He is now with the Department of Computer Engineering, North Maharashtra University, SES's R. C. Patel Institute of Technology, Shirpur, India (e-mail: sunilrathod.rathod01@gmail.com).

Tareek M. Pattewar is the Assistant Professor, He is now with the Department of Computer Engineering, North Maharashtra University, SES's R. C. Patel Institute of Technology, Shirpur, India (e-mail: tareekpattewar@gmail.com).

978-1-4799-8081-9/15/$31.00 © 2015 IEEE

experimental results shows that it has prominent performance for classification of spam image in terms of accuracy, precision, and recall [5].

Man Qi et al., has explored two main semantic methods: Bayesian algorithms and Sup-port Vector Machine (SVM). Recent spam filters are discussed in this paper for determining spam messages which utilize semantic analysis information [6].

Zhan Chuan, LU Xian-liang has presented An application to Anti-Spam Email using a new improved Bayesian-based e-mail filter. They have used vector weights for representing word frequency and adopted attribute selection based on word entropy and deduce its corresponding formula .It is proved that their filter improves total performances apparently [7].

Holly Esquivel et al. had focused on the pre-acceptance altering mechanism IP reputation. They first classify SMTP senders into three main categories: legitimate servers, end-hosts, and spam gangs, and empirically study the limits of effectiveness regarding IP reputation filtering for each category [8].

Georgios Paliouras et al., have presented Learning to Filter Spam E-Mail. They investigated the performance of two machine learning algorithm in context of anti-spam filtering by comparison of a Naïve Bayesian and a Memory-Based Approach. They have determined the performance on publicly available corpus for naive bayes. Also they have compared the performance of the Naive Bayesian filter to an alternative memory based learning approach so that in both methods accuracy has improved for spam filtering and keyword based filter are used widely for email [9].

Gray Robinson proposed computation of probability of spam mail and legitimate mail [10].

III. ALGORITHM STUDY

A. *Bayesian Classifier*

Naïve bayes classifier is popular as it is statistical classifier known for Email filtering, It uses text classification method for identifying spam mails. Naïve bayes uses tokens (words) with spam and ham mails for Calculating probability to determine whether a mail is spam or not.

Mathematical Formulation:

Bayesian classifier is based on Naïve Bayes theorem, Naïve Bayes theorem can perform more sophisticated classification methods. To demonstrate the concept consider following equations [11];

Thus, we can write:

Prior probability of Legitimate mail = Number of legitimate mail / Total number of mail. (1)

Prior probability of Spam mail = Number of spam mail / Total number of mail. (2)

Likelihood of X-mail given Legitimate = Number of legitimate mail in the vicinity of X-mails / Total number of legitimate mail. (3)

Likelihood of X-mail given Spam = Number of spam mail in the vicinity of X-mails / Total number of spam mail. (4)

Posterior probability of X-mail being legitimate = Prior probability of legitimate mail × Likelihood of X-mail given legitimate. (5)

Posterior probability of X-mail being spam = Prior probability of spam mail × Likelihood of X-mail given spam. (6)

Finally we classify X-mail as spam as its class membership has a largest posterior probability.

B. *Evaluation Criteria*

Classification results were calculated using following, As we formulate the spam detection problem as Bayesian classification problem, each mail undergo one of four possible scenarios: Accuracy (TP, correctly classified instances), Error (TN, Incorrectly classified instance), Though Error rate (fraction of wrongly classified Instances) may be of limited interest in our context where data sets are unbalanced. Additionally, we report standard measures such as precision, recall and Error.

IV. EXPERIMENT

The existing system mainly works on main headers like subject, body of mail and mailing address but we are dealing with only the body of mail which is estimated based on content. Content based filter checks for information in the body of mail by considering subjects, URLs for acceptance, rejection and classification of mail by considering content to spam and legitimate mail .The method can be described as in fig.1;

A. *Training:*

We are using mail dataset collected from Gmail which consist of spam mail and legitimate mail. This mails are considered as input in HTML format for preprocessing.

B. *Preprocessing :*

1) *HTML Tag Removal:*
The input Emails are in HTML format so this contains the tag, so to purify the text we need to remove the tags.

2) *Stopword Removal:*
This is the stopword list which consist of terms including articles, prepositions, conjunctions and certain high frequency words (such as some verbs, adverbs).

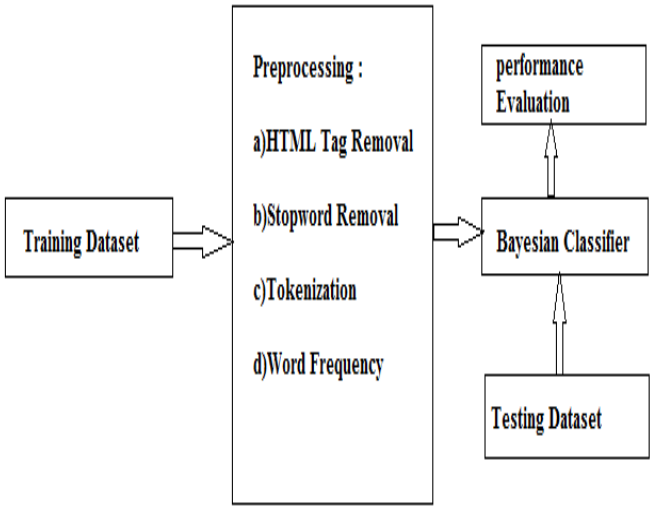

Fig. 1. Content Based Spam Detection in Email using Bayesian Classifier

3) Tokenization :

Lexical analysis also named as Tokenising, It involves dividing the content of text into strings of character called as Tokens. Filtering techniques uses white space (blank) removal and removal of punctuation symbols in tokenizing.

4) Word Frequency:

This counts the frequency of words depending on its occurrence, It helps in deriving the word probability for spam and legitimate mails.

C. Bayesian Classifier:

It is method used for classification of text, It gives efficient learning algorithm for data mining. This uses Bayes classifier theorem which is based on conditional independence assumption:

$$P(\text{spam/word}) = [P(\text{word/spam})\, P(\text{spam})] / p(\text{word})$$

Considering spam probability for words, It evaluates Spam and Legitimate mails for classification then gives performance measurement.

D. Testing Dataset :

This is derived from g-mail consisting of spam and legitimate mails .It also needs to be preprocessed to give pure text then classification is done by using Bayesian classifier. Further correctly classified instances (mails) and Incorrectly classified instances (mails) are evaluated.

E. Performance Measurement:

As classification model buids, It is essential to derive performance on the basis of parameters such as Accuracy (Correctly classified instances), Error (Incorrectly classified instances), precision and Recall are evaluated .

$$\text{Accuracy} = (TN + TP) / (TN + TP + FN + FP)$$

$$\text{Precision} = (TP) / (TP+FP)$$

$$\text{Recall} = (TP) / (TP + FN)$$

Where,
TN: True Negative, Legitimate predicted as Legitimate
TP: True Positive, Spam predicted as Spam
FP: Legitimate predicted as Spam
FN: Spam predicted as Legitimate.

V. RESULT

A. Computation of Bayesian classifier efficiency under different data volume:

We are performing experiments on different volume of training dataset and testing data set form Gmail .
The training and testing dataset of Gmail are 1000 mails, 1500 mails and 2100 mails then we are going to measure the performance in terms of Accuracy, Error, Time, Precision and recall. The fig.2 describes Accuracy for the system architecture defined in fig.1 under different volume of dataset. Similarly Error rate can be shown in fig. 3, whereas the time needed to perform this classification and filtering can be deduced with fig.4. The Precision and Recall can be shown in fig.5 and fig.6 respectively. Finally Performance Measurement can be shown with Table I.

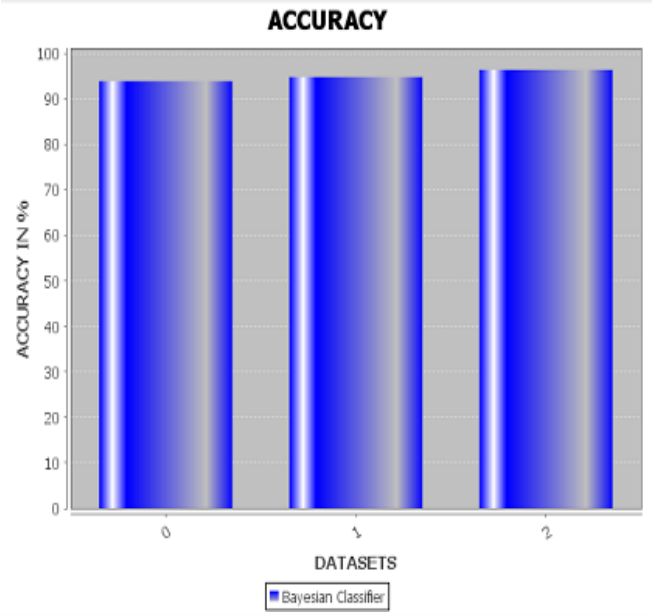

Fig. 2. Derived Accuracy on different volume of dataset

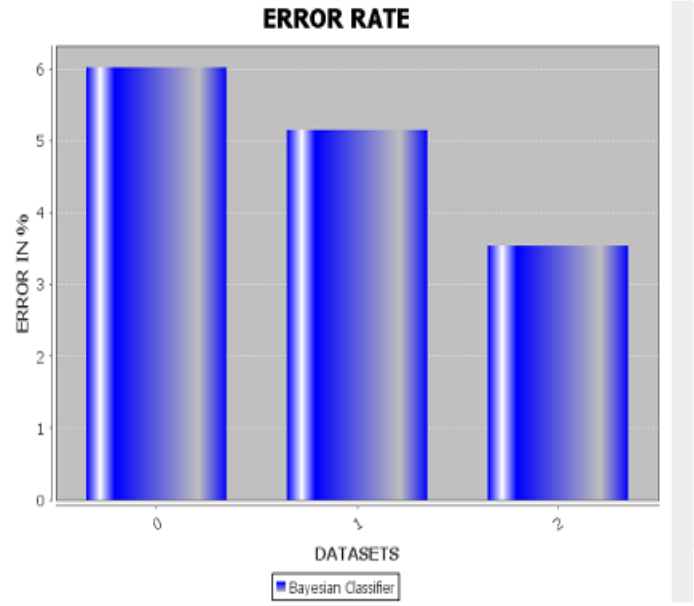

Fig. 3. Derived Error Rate on different volume of dataset

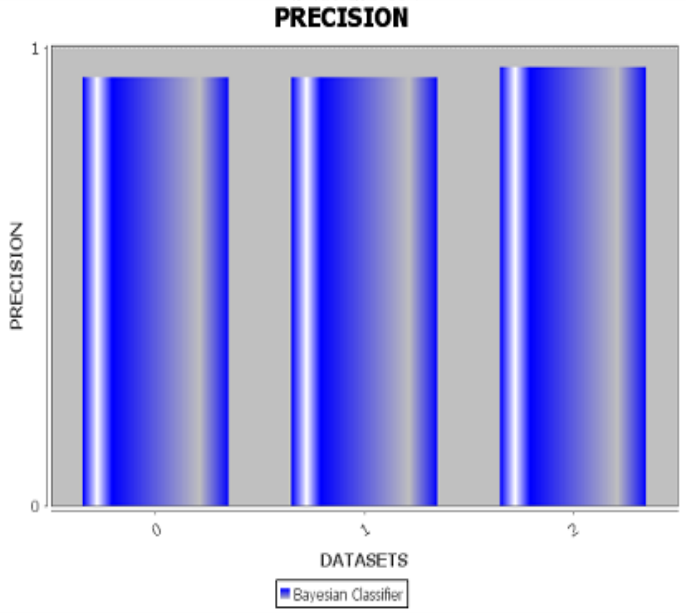

Fig. 5. Precision on different volume of dataset

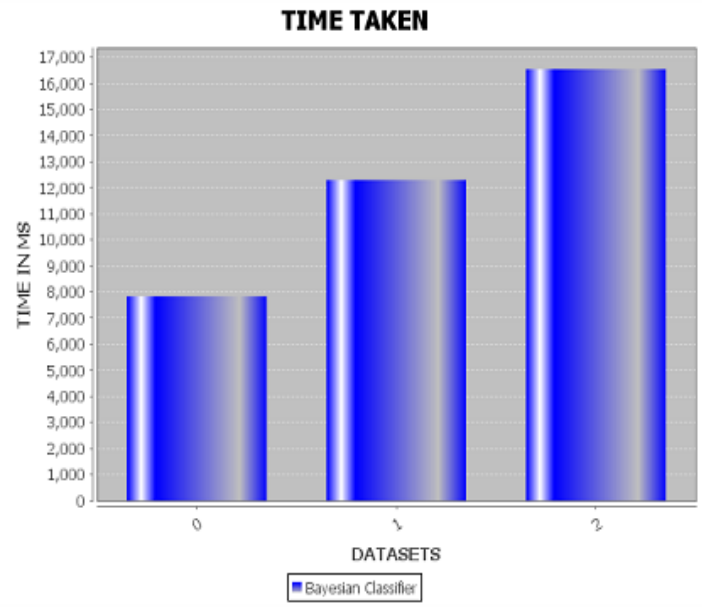

Fig. 4. Time taken on different volume of dataset

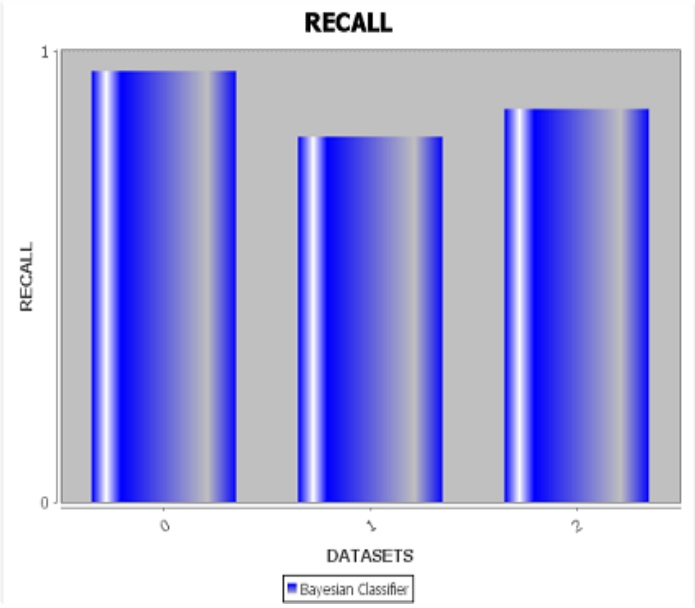

Fig. 6. Recall on different volume of dataset

TABLE I
PERFORMANCE MEASUREMENT

Bayesian Classifier	(TP) Accuracy (%)	Error (TN) (%)	Time (MS)	Precision	Recall
DATASET 1	93.98	6.02	7834.0	0.93	0.95
DATASET2	94.85	5.15	12294.0	0.93	0.81
DATASET3	96.46	3.54	16546.0	0.95	0.87

VI. CONCLUSION

We have emphasized Bayesian approach for classifying Spam and legitimate mails using supervised learning across features extracted. Applying the Bayesian classifier, we experimentally demonstrated that spam mails can be detected with an accuracy of more than 96.46% with respect to real-world gmail data sets. The mail dataset once trained, effectively detect a potentially spam mails and thus help internet users from avoiding those spam.

As future work, We will integrate these content based spam detection System with malicious URL detection to improve the accuracy of the system for detecting spam mails and malicious URLs.

ACKNOWLEDGMENT

We are sincerely grateful to all the persons who help us through this work to make it successful.

REFERENCES

[1] Dhanalakshmi Ranganayakulu and Chellappan C., "Detecting malicious URLs in E-Mail - An implementation", *AASRI Conference on Intelligent Systems and Control*, Vol. 4 , 2013,pg. 125–131

[2] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail", *AAAI Tech. Rep.WS-98-05*, pp.55–62, 1998.

[3] V Christina., "A study on email spam filtering techniques", *International Journal of Computer Applications*, Vol. 12– No.1, 2010.

[4] Sadeghian, A and Ariaeinejad, R., "Spam detection system: A new approach based on interval type-2 fuzzy sets", *IEEE CCECE -000379*, 2011.

[5] Congfu Xu, Yafang Chen, Kevin Chiew, "An approach to image spam filtering based on base64 encoding and N-Gram feature extraction", *IEEE International Conference on Tools with Artificial Intelligence*, DOI 10.1109/ICTAI.2010.31, 2010.

[6] Man Qi, Mousoli, R,"Semantic analysis for spam filtering", *International Conference on Fuzzy Systems and Knowledge Discovery*, Vol.6,Pg.2914-2917, 2010.

[7] Zhan Chuan, LU Xian-liang, ZHOU Xu, HOU Meng-shu, "An Improved Bayesian with Application to Anti-Spam Email ", *Journal of Electronic Science and Technology of China*, Mar. 2005, Vol.3 No.1

[8] Holly Esquivel and Aditya Akella, "On the effectiveness of IP reputation for spam filtering", *IEEE International Conference on Communication Systems and Networks*, DOI:10.1109/COMSNETS.2010.5431981, Pg.1-10, 2010.

[9] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memorybased approach", *Proceedings of the Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000)*, pages 1–13, 2000.

[10] G. Robinson. (2014,Oct). "A statistical approach to the spam problem", 2003. [Online] Available : http://www.linuxjournal.com/article.php?sid=6467

[11] Naïve Bayes Classifier.(2014, Dec) [online] Available : http://www.statsoft.com/textbook/naive-bayes-classifier