INTEGRATED SPAM DETECTION FOR MULTILINGUAL EMAILS

Akash Iyengar[1] , G.Kalpana[2], Kalyankumar.S[3] ,S.GunaNandhini[4]
akashiyengar72@gmail.com, g.gkalpana@gmail.com, kalyanshaddy@gmail.com,
gunanandhinis@saec.ac.in

Abstract: - Emails and social communication is the newest and easiest way of data communication. Although Electronic communications are facile in nature, it is equally easy to attack these services with an intention of fraud and trickery motivation. The cyberpunks use emails to obtain valuable credentials which could be used in many ways. A simple looking mail can include phishing URLs for purloin of important information. To avoid such criminal activities, perceive and filter such kinds of mails, an integrated approach is used. Each email provider has their own filtering system or technique, but most of them don't work to their full capacity. There are instances where an email in native language or any other language other than English can be considered genuine. The objective of the paper is to overcome this problem by using an integrated technique to increase the efficiency of detecting and filtering the emails.

Key terms: - Spam filtering, Bayesian filtering, Greylist filtering, Email Classification

I. INTRODUCTION

Electronic mail services are the most efficient, time saving and easiest means of information communication. The simplicity of email communication has also generated troubles to its users through junk and phishing mails. These mails are also known as spam mails. Usually spam mails are sent in bulk to many users simultaneously. They could be generic mails or mails for advertising purposes. There is no definite way to determine which mails is genuine and which mails is malicious. There are mails which contain genuine advertising information but at the same time they might also contain malicious URLs, these mails are commonly known as phishing mails. Spreading virus is another important aspect of spam emails. These kinds of services are in common practice by hackers and phishers for extortion of money and for damaging social reputation. To overcome such drastic problems one should be taught not to reply to such mails or click the URLs provided in such emails.

Integrated Spam Filter:

Integrated spam filter is a combination of both content based filter as well as List based filters. Content based filters are used to check structural contents of the emails. It mainly focuses on body of the mails and the URLs provided in the body as a text. It also takes into account the headers or the subject of the mail. It is based on classification of text by implementing preprocessing of text in terms of HTML tag removal, Tokenizing , Frequencies of words and stop word removal to find out if a given mail is a spam or not. It also uses List based filters for stopping spam by categorizing senders as spammers or trusted sources and blocking or allowing their messages accordingly.

The existing system filters mails using only content based filters, which is not as efficient in tracking the URLs through which these mails came. Based on the previous research, it is estimated to have an efficiency of 96.6%.

The Whole paper is organized into six sections. The following section would describe the related work on filtering methods and spam classification. The third section describes the algorithm study. We have used Grey list algorithm and Bayesian classifier algorithm along with evaluation criteria for our work. The fourth section explains the experiments done for the research, which includes training sets, preprocessing of data, application of both Bayesian and Grey list classifiers and testing of data sets. The fifth section elaborates the results of the experiments with due consideration of performance measurement parameters. The sixth and final section would conclude the paper.

II. RELATED WORK

The existing work underwent an implementation on detection and filtering of emails and malicious URLs in Emails using Bayesian Classifiers by Sunil B. Rathod, Tareek M. Pattewar 2016, they had considered only content based filtering technique for classification and filtering of spam emails [1].

Implementation on detection of malicious URLs in Email by Dhanalakshmi Rand Chellapan C 2013, they had considered Age of Domain along with host based and lexical features. They had also used page ranking system for analysis of URL.They had also used Bayesian classifiers to improve accuracy for reduced data sets [2].

978-1-5090-6135-8/17/$31.00 ©2017 IEEE

Sahami et al 1998 has provided with a spam classification technique using Bayesian approach. Bayesian Approach is considered the most advanced way of classification and filtering technique. It is a statistical classifier which is based on probability. They have considered contents of emails with features of domain and visualized that accuracy could be improved [3]. V Christina et ai, had shown the need of effective spam filters has increased. He had discussed spam and it's filtering Methods along with their correlated problems [4].

Sadeghian A. et ai, had exploited spam detection based on interval type-2 fuzzy sets. It provides user more control on categories of spam and permits the personalization of the spam filter [5].

Zhan Chuan, LV Xian-liang has presented an application to Anti-Spam Email using a new improved Bayesian-based email filter. They had used vector weights for representing word frequency along attribute selection based on word entropy and deduce its corresponding formula .It is proved that their filter had provided with improved results [6].

III. ALGORITHM

STUDY *A. Bayesian Classifier*

It is a statistical classifier known for filtering emails. It uses text classification for the filtering purpose. Naïve Bayes uses tokens or words with spam and ham mails for crafting probability to check if a mail is a spam or genuine.
Mathematical Formulation:

It is based on Naïve Bayes theorem. It can perform high level classification. The following equations can be used to demonstrate the concept of Naïve Bayes [7].
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. (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 x Likelihood of X-mail given legitimate. (5)

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

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

B. List Formation

The list of spammers is usually formed for protecting mails from spamming. The list can be created by an individual who can have a definite set of spammers .The user can either block a list of users or allow a list of users. Third party organizations can also be used for creating lists. They create list on a larger scale almost covering all the users who have been reported as spammers by individual users.

C. Evaluation Criteria

We can formulate the spam detection problem as Bayesian classification problem,

Each mail can be classified under any of the four scenarios: Error (TN, Incorrectly classified instance), Accuracy (TP, Correctly classified Instance). Error rate is not of much interest in our context where data sets are usually unbalanced. We also report precision and recall.

IV. EXPERIMENT

The current system mainly gives emphasis on main headers, subject line, body of the mails and the URLs in the body of the mail, but we are considering both body of the mail along with the predetermined list of spammers provided by third party organizations. The content based filters checks for the URLs in the body of the email whereas the List based filter reject the mail and shows failure message to the sender if found as spam. The method can be described as in fig 1:

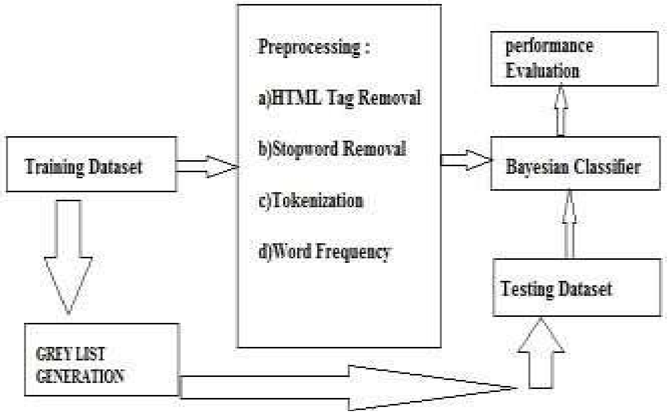


Figure 1: Integrated Spam Detection In Email Using Bayesian and Grey list

A. Training

We are using mail data sets collected from Gmail and Yahoo which consists of spam as well as genuine mails. They are considered as an input in HTML format for preprocessing. It also has the Preexisting spammers list provided to Gmail or Yahoo.

B. Preprocessing

1) Tag Removal

As input is in the form of HTML tags, these tags have to be removed in order to process the data.

2) Stop word Removal

It includes the removal of predefined words, prepositions, verbs, adverbs, nouns and adjectives.

3) Tokenization:

Lexical analysis is done on the data by dividing the content of the text into characters. It also removes white spaces and punctuations.

4) Frequency of words

These counts the frequency of words based on its occurrence. It is used for determining the probability for spam mails

C. Bayesian Classifier

It is an efficient learning algorithm for classification of text in data mining. It is based on independence assumption:

P (spam/word) = [P (word/spam) P (spam)] / p (word)

Considering spam probability for words, it evaluates Spam and Legitimate

Mails for classification then give performance measurement.

D. Greylist Filtration

It has a list of potential spammers, who have been flagged by the users of Gmail or other mail services. As soon as the system detects flagged users, it rejects and mail. If the mail is sent by a legitimate user then the user would send the mail again after reading the failure message by their server. It basically increases the efficiency of filtering process.

E. Testing Dataset:

The spam as well as legitimate mails are received from Gmail or Yahoo as an input and are classified as correctly classified instances or mails and incorrectly classified instances or mails. This classification is done through the Bayesian Classifiers and is evaluated.

F. Performance Measurement:

Performance based on the parameters like Accuracy, precision, Recall and Error are evaluated. They can be formulated as below:

Accuracy = (TN + TP) / (TN + TP + FN + FP)
Precision = (TP) / (TP+FP)
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. Classification of Efficiency under different data volume:

The experiment is done on different volumes of training data set in comparison to the testing data set from Gmail or Yahoo. The volumes of datasets are 1200 mails, 1600 mails, 2200 mails. We have measured 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.

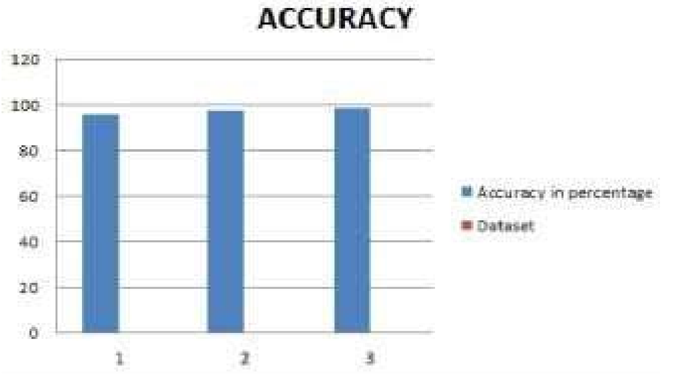
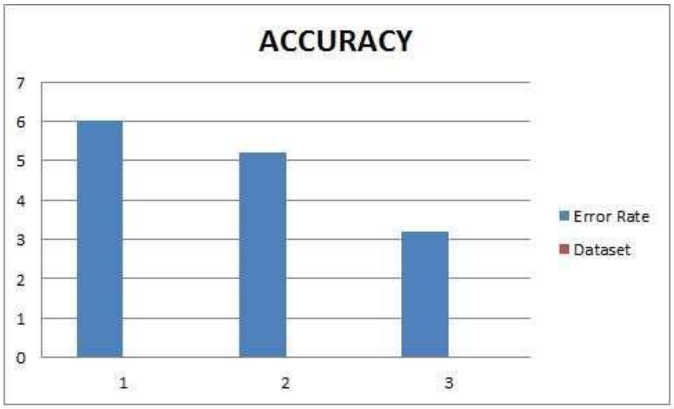


Figure 2: Derived Accuracy on different volume of dataset



. Figure 3: Derived Error Rate on different volume of dataset

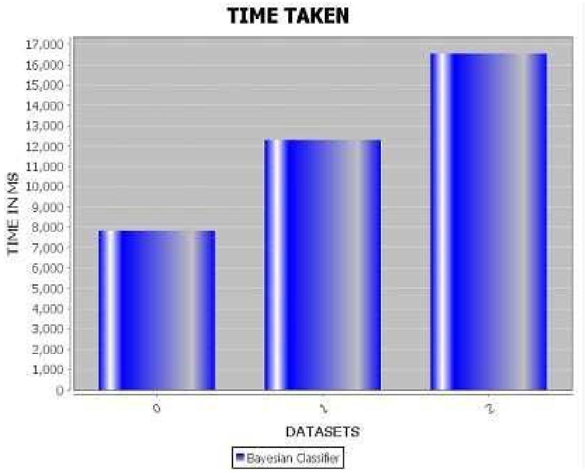


Figure 4: Time taken on different volume of dataset

978-1-5090-6135-8/17/$31.00 ©2017 IEEE

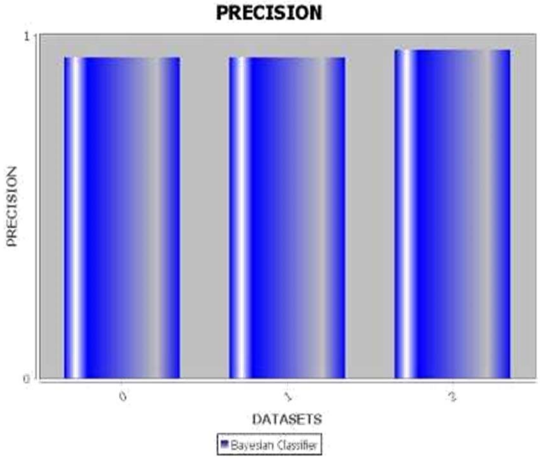


Figure 5: Precision on different volume of dataset

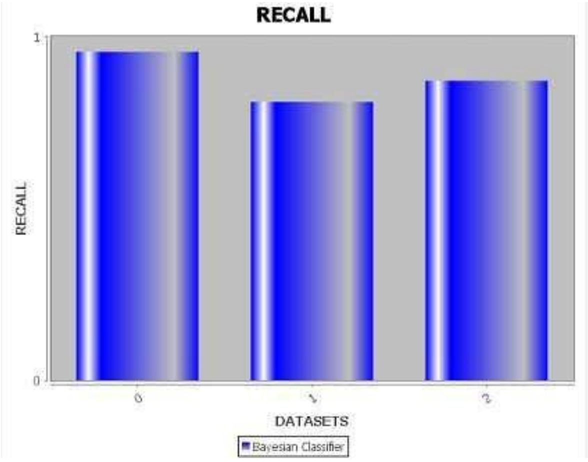


Figure 6: Recall on different volume of dataset

TABLE 1: Performance Measurement

Bayesian Classifier	(TP) ACCURACY (%)	ERROR (TN) (%)	TIME (MS)	PRECISION	RECALL
DATA SET 1	95.98	6.00	7835.0	0.94	0.95
DATA SET 2	96.66	5.1	12294.0	0.93	0.82
DATA SET 3	97.3	3.3	16546.0	0.96	0.88

VI. Conclusion

We have emphasized on an integrated approach for classifying and filtering of spam and legitimate mails. Applying the integrated approach over the traditional approach has increased the accuracy by more than 1% from 96.46% to 97.3% with respect to the real world data set. It basically helps internet users to avoid spam mails.

As future work we will try to increase the accuracy of spam detection for the users dynamically in real time by integrating the current approach with the URL detection framework.

Acknowledgment

We would like to thank all the people who have helped us throughout the work for making it successful.

References

[1] Sunil B.Rathod, Tareek.M.Pattewar “Content Based Spam Detection in Email using Bayesian Classifier”,ICCSP Conference ,2015.

[2] Dhanalakshmi Ranganayakulu and Chellappan C., "Detecting malicious URLs in E-Mail - An implementation", AASRl Conference on intelligent Systems and Control, Vol. 4, 2013.

[3] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail ... AAAiTech.Rep.WS-98-05. pp.55-\62, 1998.

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

[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,2010.

[6] Zhan Chuan, LU Xian-Iiang, 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.

978-1-5090-6135-8/17/$31.00 ©2017 IEEE