Spam Email Detection using Structural Features

Sarju S
PG Scholar
Department of Computer Science & Engineering
KCG College of Technology, Chennai-97

Riju Thomas
PG Scholar
Department of Computer Science & Engineering
KCG College of Technology, Chennai-97

Emilin Shyni C
Associate Professor
Department of Computer Science & Engineering
KCG College of Technology, Chennai-97

ABSTRACT
In recent years, we have witnessed a dramatic raise in the use of web and thus email becomes an inevitable mode of communication. This is the scenario where the attackers take advantage by the mode of spam mails to the email users and misguide them to some phished sites or the users unwittingly install some malwares to their machine. This shows the importance of research activities being carried out in the field of spam mail detection. In this paper we tend to project a replacement methodology to segregate spam emails from non-spam (legitimate) emails using the distinct structural features available in them. The experiments with 8000 emails show that that our methodology preserves an accuracy of the spam detection up to 99.4% with at the most 0.6 % false positives.

Keywords
Spam Detection; Structural Feature Selection; spam classification;Machine learning application.

1. INTRODUCTION
An aggressive innovation that has grabbed the globe nowadays with its magic speed in communication is e-mail. However several challenges are being faced by this service like email worms, spam emails and phishing emails, out of that most distinguished type of email attack is spamming. This process highlights passing similar message to several users. Spam mails clearly called junk emails received from anonymous sources contains invalid or pretend data and should have malwares hooked up thereto or URLs that misguide the user to some phishing sites. An email is claimed to be a spam if its supply is unknown or belongs to mass mailing and also the email response is not requested by the recipient. According to Spam Trends and Statistics Report 2013 by Kaspersky Lab [9] the percentage of spam increases in the total email traffic to an alarming rate of 70.17%, also the emails with malicious attachment also increases. The studies unveils a brand new trick of spammers that style mails with malicious attachment that precisely seem like automatic delivery failure notifications sent by the e-mail servers or seem like notifications from well-known social networking sites.

In order to detect spam mails many techniques are enforced. Classification algorithms based mostly systems offers higher results among these projected works. Classifiers implements algorithm like Naïve Bayes, Decision Tree, and Support Vector Machines (SVM) that uses the dataset collected from the emails to tell apart ham (legitimate mails) from spam mails. Classifiers works by relating the utilization of feature set (usually words) collected from dataset then uses the corresponding algorithm to compute the chance that associate email is spam. During this paper we tend to conduct experiment on information sets containing 8000 emails and structural properties are collected that is employed for detecting whether or not the e-mail is spam or ham. The accuracy of the spam detection is evaluated using the machine learning algorithms Bayes Naïve Thomas Bayes, Random Forest and AdaBoost.

2. BACKGROUND
Spam Mails are one among the foremost common and high problem within the internet which may bring large damage to organizations similarly as for individual internet user. Spam email additionally referred to as junk email or unsought bulk email that is causing identical messages to various recipients. It takes plenty of time to spot, arrange and delete these mails. There can also be an opportunity of counting a legitimate mail as a spam mail. To handle this threat, several techniques are introduced for detecting spam mails that are mainly based on artificial intelligence and data mining approaches. Data mining based techniques are mostly used as a result of their less complexity to implement and provides higher results than the artificial intelligence based ones. In Carreras et al. [4] they have used Adaboost classifier for filtering spam emails and discerned that on increasing the complexity of base learners high exactitude can be achieved. But the main setback of this methodology is that as the spammers change their strategies oftentimes, it becomes terribly costly to calculate and recalculate using this technique. Sahami et al. [12] used Bayesian analysis, to calculate the likelihood that an email with a particular set of words in it belongs to either ham or spam class. If the email's spam likelihood computed exceeds some bound threshold, the filter can mark the e-mail as a spam.

In several email data sets, solely a little feature of the entire features collected is helpful for categorizing ham or spam classes. Shankar et al. [13] developed an algorithm to learn weights for the various features collected that solve a significant downside of filters, that they are not using all the collected features for computing. This approach extensively will increase the classification accuracy. The classification errors will be brought down by discovering temporal relations in an email within the type of temporal sequence pattern as proposed by Kiritchenko et al.[10].

Spam filtering techniques are mainly based on text categorization methods. Thus filtering spam is a type of classification problem. In this paper, the structural properties that are commonly available in the email isextracted and classified as spam or ham using classification methods. The machine learning algorithms used in this paper are Naïve Bayes, Random Forest and AdaBoost.

2.1 Naïve Bayes

Naïve Bayes is a statistical classifier that calculates the posterior probability of each class of the target attribute (in this case, spam or ham) based on the values of training data so that a given email can be assigned to the class with highest probability. Naïve Bayes classifier is based on Bayes theorem [14]. From the theorem we can state that the probability of an email message with structural set vector set $x=<x_1 \ldots x_m>$ belongs to a category C (either ham or spam) is

$$p(c|x) = \frac{p(c)p(x|c)}{p(x)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

2.2 Random Forest

Random forest classification is introduced by Breiman[3], builds a randomized decision tree during each iteration of the algorithm. It also has an effective method to handle the missing data and its precision is very good. To classify a new object from given structural vector set, use the vector down each of the trees in the forest. Each tree in the forest indicates a classification. The forest selects the classification having the most number of votes.

2.3 AdaBoost

AdaBoost is the most popular machine learning algorithm introduced by Freund et al.[8].The reason for the wide acceptance of AdaBoost algorithm are , they are very sensitive to noisy data and less susceptible to the over fitting problem. The AdaBoost is flexible to combine with any other learning algorithms to improve the performance of classification. The AdaBoost algorithm is shown in the Figure 1.

1. Start with initializing weights $w_j=1/N$, $j=1,2,\ldots N, F(m)=0$
2. Repeat for $k=1,2,\ldots K$
   a. Fit the regression function $f_k(x)$ by weighted least squares of y_i to x_i with weights w_j.
3. Recalculate $F(m) = F(m) + f_k(x)$
4. Update weight using the equation $w_i = w_i e^{-y_i f_k(x_i)}$ and renormalize.

Figure 1 -AdaBoost Algorithm

3. PROPOSED SOLUTION

In this paper we propose a new methodology for detecting spam emails based on the structural properties of the email. The structural properties of the email are analyzed to separate the spam emails from the ham. Publically available spam email data set are used as the input to the proposed methodology. Figure 2 shows the proposed methodology.

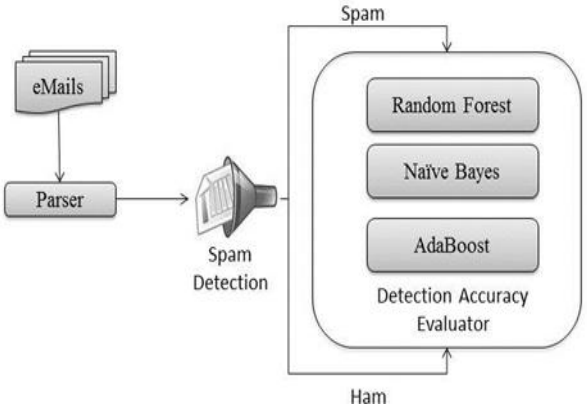

Figure 2 – Proposed Solution

3.1 Parser

Raw emails data are typically present in Multipart Internet Mail Extension (MIME) format. The parser parses the email document and extracts body text, hyperlinks and words present in the body of the email. The Apache James Mime4[2]is used in the development of the parser for extracting content from e-mail message streams in plain Multipart Internet Mail Extension (MIME) format. The parser uses a call-back mechanism to report parsing events such as start of an entity header, the start of a body, among others. It solely deals with the structure of the message stream and has been designed to be extremely tolerant against messages violating the standards. Structural features are extracted from the parsed document.

3.2 Spam Detection

Spam Detection is done using the structural features present in the email. In this paper 46 structural features present in an email are used. The Table 1 shows the structural features extracted. Algorithm for the spam detection is shown in the Figure 3. The variable *V* represents the feature vector,

1) Create the feature vector *V* as a string of length N, where each bit in it represents a particular spam feature which used for spam detection
2) Generate the feature vector bit values
3) For J=1 to M
   a. α = randomize()
   b. Detect the class label S using the α
4) End For

Figure 3 – Spam Detection Algorithm

N indicates the length of the feature vector, *M* represents the total number of feature sets, α indicates the randomly selected spam feature.

3.3 Detection Accuracy Evaluator

The final stage of the proposed methodology is the measurement of the detection accuracy. Three classifiers are used here for evaluating the accuracy of spam detection. The different measures used for evaluating the performance are True Positive Rate, False Positive Rate, Precision, Recall and F-Measure. The dataset contains different combinations of spam and ham mails.

Table 1- Structural Features

No	Feature	No	Feature
1	bodydearword	24	scriptonclick
2	bodyform	25	scriptpopup
3	bodyhtml	26	scriptstatuschange
4	bodymultipart	27	scriptunmodalload
5	bodynumchars	28	senddiffreplyto
6	bodynumfunctionwords	29	sendnumwords
7	bodynumuniqwords	30	sendunmodaldomain
8	bodynumwords	31	subjectbankword
9	bodyrichness	32	subjectdebitword
10	bodysuspensionword	33	subjectfwdword
11	bodyverifyyouraccountphrase	34	subjectnumchars
12	externalsabinary	35	subjectnumwords
13	externalsascore	36	subjectreplyword
14	scriptjavascript	37	subjectrichness
15	subjectverifyword	38	urlatchar
16	urlbaglink	39	urlip
17	urlnumdomains	40	urlnumexternallink
18	urlnumimagelink	41	urlnuminternallink
19	urlnumip	42	urlnumlink
20	urlnumperiods	43	urlnumport
21	urlport	44	urltwodoains
22	urlunmodalbaglink	45	urlwordclicklink
23	urlwordherelink	46	urlwordloginlink

4. EXPERIMENTS

The experiment is conducted using the various data sets and results are reported. The data sets 1, 2, 3 and 4 contain 2000, 4000, 6000 and 8000 emails respectively. True Positive (TP) means that the particular and the actual classes are positive and False Positive (FP) means that the expected ought to have the negative, classified instead as positive. Alternative performance metrics employed in classifications are accuracy, precision, recall and F-measure. K fold cross validation is employed within the classifier to judge the performance. The training set is randomly partitioned into k disjoint sets. If the value of the k set to 10, then it becomes 10 fold cross validation. In 10 fold cross validation 90% of the given data is employed to train the classifier and remaining 10% data are used as test data.

One of the ways to visualize the performance of the classifier algorithm is the confusion matrix. In which every row has actual values and every column has expected values. Confusion matrix using dataset 1 is shown in the table 2.

Performance of the different classifiers is visualized in the confusion matrix.

Figure 4 shows the precision comparison of spam mails on totally different data sizes. The precision is calculated by using the equation 2. The result shows that performance of the Random Forest is good comparing to other classifiers on all data sizes.

$$Precision = \frac{True\ \ Positive}{True\ \ Positive\ \ + False\ \ Positive} \quad \ldots\ldots\ldots (2)$$

Table 2– Confusion matrix

Random forest		
	Spam (predicted)	**Ham (predicted)**
Spam (actual)	4097	19
Ham (actual)	30	4120
Overall accuracy		99.41%
Adaboost		
	Spam (predicted)	**Ham (predicted)**
Spam (actual)	4019	97
Ham (actual)	49	4101
Overall accuracy		98.23%
Naïve bayes		
	Spam (predicted)	**Ham (predicted)**
Spam (actual)	4068	48
Ham (actual)	34	4116
Overall accuracy		99%

Table 3 shows the recall of different classifier once different datasets are applied. The result shows that the performance of the Random Forest is nearly same on all data sizes. Performance of the Naïve Bayes shows immense variations and Adaboost is additionally having some variations within the preciseness. Figure 5 shows the accuracy of the spam detection. From the results it's understood that detection of spam mail is pretty sensible once data size is of 8000 mails.

Table 3-Recall comparison of spam mail detection

Data Size	Random Forest	Naïve Bayes	AdaBoost
2000	97.7	71.9	91.5
4000	95.3	64.8	89.8
6000	98.6	97.7	98.1
8000	99.1	99	98.6

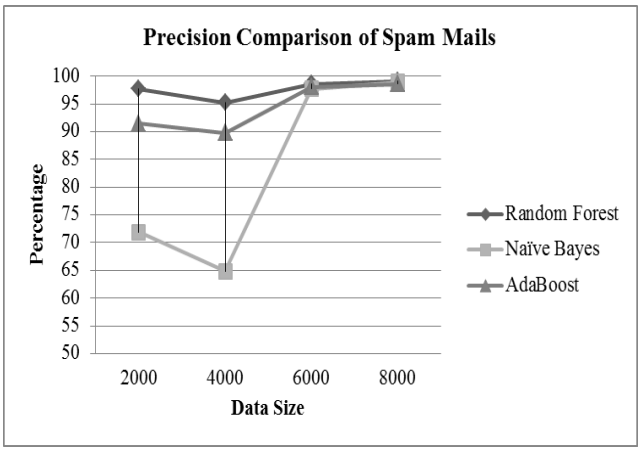

Figure 4- Spam mail detection precision comparison

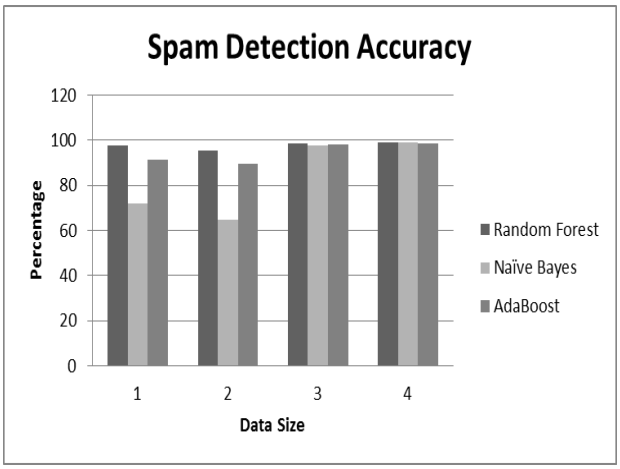

Figure 5- Comparing accuracy of spam detection

5. CONCLUSION

The proposed methodology demonstrates the flexibility to detect spam via appropriate identification and usage of structural properties of email. The feature vector includes forty six structural features from the header and body part of the e-mail. Our argument is that single data set wouldnot be adequate to urge the clearest image of the accuracy, so the experiments done on numerous data sets. It is obvious from our experimental results that various classifier offers various results on the different data sets. The results show that our methodology preserves an accuracy of the spam detection up to 99.4% with at the most 0.6 % false positives.

In this work we tend to think about solely mails which will be parsed using MIME parser. Some spam mails could associate with the photographs and attachments are unable to parsewithin the current work.Future work ought to expand this design and modify it to spot all kinds of emails.

6. REFERENCES

[1] S. Abu-Nimeh, D. Nappa, X Wang, S. Nair, "A comparison of machine learning techniques for phishing detection." In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, pp. 60-69. ACM, 2007.

[2] Apache James Mime4, http://james.apache.org/mime4j/

[3] L .Breiman, "Random Forests.," InMachine Learning,Vol 45 No.1,pp.5–32,2001.

[4] X.Carreras,L.Marquez and J.G Salgado ,"Boosting trees for anti spam filtering," In International conference on Recent Advances in Natural Language Processing, , 2001,pp.58-64.

[5] J Clark, I Koprinska, J Poon," A neural network based approach to automated e-mail classification," In Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on, pp. 702-705. 2003.

[6] I Fette, N Sadeh, A Tomasic, "Learning to detect phishing emails." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[7] G .Forman, "An extensive empirical study of feature selection metrics for text classification.,"InThe Journal of machine learning research,pp.1289-1305,2003.

[8] Y.Freund , R. E. Schapire, "A Short Introduction to Boosting.,"1999.

[9] Kaspersky Spam Trends and Statistics Report (2013), http://www.securelist.com/en/analysis/204792297/Spam_in_Q2_2013

[10] S.Kiritchenko,S.Matwin,SAbu-Hakima."Email Classification with Temporal Features," In Intelligent Information Systems,2004,pp.523-533

[11] M .Rathi, V. Pareek, "Spam Mail Detection through Data Mining-A Comparative Performance Analysis.," InInternational Journal of Modern Education & Computer ScienceVol 5 No.12,2013.

[12] M.Sahami, S.Dumasi, D.Heckerman, and E.Horvitz, "A Bayesian approach to filtering junk e- mail: In Learning for text categorization," InInternational Journal of Modern Education and Computer Science (IJMECS), Vol.5 No.12,pp.31-39,1998.

[13] S. Shankar and G. Karypis, "Weight adjustment schemes for a centroid based classifier," Computer Science Technical Report TR00-035, 2000.

[14] B.Thomas, and P.Richard, "An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S.," InPhilosophical Transactions of the Royal Society of LondonVol 53 No.0,pp.370–418,1763.

[15] L Zhang, J Zhu, T Yao," An evaluation of statistical spam filtering techniques.," In ACM Transactions on Asian Language Information Processing (TALIP) Vol 3, No. 4,pp. 243-269,2004.