International Journal of LNCT, Vol 2(5)
ISSN (Online): 2456-9895

Review of spam classification using different machine learning algorithms

Aditya Shrivastava and Rachana Dubey
LNCT, Bhopal

Abstract
Email is necessary and essential for communication in today's life. Today internet users are increases, and email is necessary for communication over the internet. Spam mail is a major and big problem of researchers to analyze and reduce it. Spam emails are received in bulk amount and it contains trojans, viruses, malware and causes phishing attacks. Problems are arise when number of unwanted mails are come from unknown sites and how to classify the user that email are received which is spam email or ham. This paper used to classify that incoming emails are spam mail or ham by the use of different classification techniques to identify spam mail and remove it. This paper benchmark dataset is used. The dataset contains 58 attributes and 4601 instances used to build a model. These papers play a very important role to remove viruses, trojans, malware and websites including phishing attacks and fraudulent attempts in emails. Different Classification algorithms are used like Naive Baye's, Random Forest and Random Tree are applied on spam dataset.

Keywords
Spam problem, Anti spam , Classification methods.

1.Introduction
Email is mandatory for every field of communication in modern era such as Education, Banking, Social networking sites, and offices etc. Spam mail means unwanted mail or unused mail, that means there are no use in present and not in future. Ham is opposite for spam means wanted and useful mail. To detect incoming mail is spam or ham is a big problem for users. Every user tries to recognize spam mail and to delete it. But the spam mail is coming in huge amount, it contains advertisement of some commercial websites for purchasing their products, also contains fraud messages to open free saving account and apply for credit card and to credential information of user causes phishing attacks. So to remove the problem researchers are used different spam filtering approaches and machine learning algorithms to identify spam messages.

Why people send spam
Spam mail is equivalent to junk mail. Spammers send spam to promote business by advertising of selling products online shopping websites. Also send spam messages to hack details of user's personal confidential information by using fake websites appending bank links. Also spam mails to destroy users system by appending Trojans, viruses, and malwares [1].

Problems with spam mail
Cost - Costs to handle spam is very high because spam mails will come in bulk amount in each and every account of users and these causes wastage of internet data and bandwidth. It is key factor to calculate total costs to handle spam at global level[1].

Privacy - spammers send fake websites link bank websites to login with password and user id or apply for credit card to capture user credential information and that misuse the information fraud attempts to use users account[1].

Time - time is the key factor in life because no one have time to check the email is coming my mail is spam or not and not sufficient time to check headers information in which id mail will come , in busy life schedule spam mail is the major problem to handle with sufficient time because it comes huge amount[1].

Security - in number of huge spam mails come in everyday, security of the system violates because deception and fraud attempts to capture users credential information i.e. to open a free account by using name, mobile no., email, and aadhar no. And personal details like debit card [1].

2.Working of anti spam
It completes in two parts are action and classification. In classification phase, the message is subject to define it is spam or not. It is also used to define the message is spam or ham and looking its characteristics to derive it[1].

In action phase, once the classification is done the message is spam or not, it will be rejected or selected for move into mailbox [1].

Step-1: collect the spam and non spam data
Step-2: Preprocessing the data means to remove the unused field and refresh the data.
Step-3: classify the bad sender specific or check the header of mail.

Step-4: apply appropriate classification methods to test according of our requirement.
Step-5: Result stored in text based, image based or content based[1].

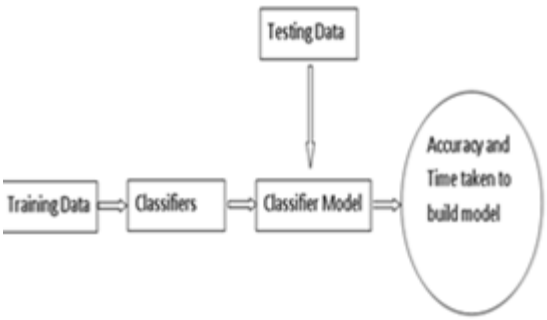

Figure 1 How anti-spam words

3.Literature review

In 2015, Anju Radhakrishnan and Vaidhehi performed a work, "Email Classification Using Machine Learning Algorithms". They used two machine learning algorithms namely j48 classifier and naïve baye's classifier. Email classification Weka software is used. And gives result that j48 classifier is more efficient than naïve baye's classifier [2].

In 2016, Prof Manisha Tijare, Mis.ElifeneshYitagesu, performed a work, "Email Classification using Classification Method". They used three classification algorithms namely support vector machine, k nearest neighbour method and naïve baye's classifier. In this classification Weka software is used. And gives result that naïve baye's classifier is more effective than other [3].

In 2015, Mr. C. Balakumar, Dr. D. Ganeshkumar performed a work" Data Mining on Various Classifiers in Email Spam Filtering". They used six classification algorithms namely BFTree, REPTree, Random Forest Tree (RndTree), SimpleCARTAlgorithm, Logistic Model Tree Induction, J48 Decision Tree Algorithm. In this classification Weka Software is used. The Random Tree gives best performance than others [4].

In 2017, Shahreen Kasim, Norfaradilla Wahid, Nurul Fitriah Rusland, HanayantiHafit performed a work, "Analysis of Naive Bayes for Spam mail across Multiple Datasets". They used only naïve baye's classifier and with weka software. And gives result that naïve baye's perform a role in spam filtering algorithms [5].

In 2015, Priyanka Sao, Prof. KarePrashanthi performed a work," Spam mail Classification by Naïve Bayesian Classifier". They used only naïve baye's classifier and with weka software. And gives result that spam mail is big problem and can be solve by using naïve baye's classifier [6-15].

4.Proposed work

Table 1.1 Evaluation measures for spam filters

Evaluation Measure	Evaluation Function
Accuracy	$Acc = \frac{TN + TP}{TP+FN+FP+TN}$
Recall	$r = \frac{TP}{TP+FN}$
Precision	$P = \frac{TP}{TP+FP}$
F-measure	$F = \frac{2pr}{p+r}$

Evaluation measures for spam filters are the testing process of dataset [5].
Accuracy and time is the main factor to compute and build the model.

Where, precision = ratio of tp to false positive, true positive.

Recall = ratio of tp to the false negative, true positive.
Accuracy= ratio of sum of true negative and true positive to the sum of all number of instances such as false positive ,true positive, false negative,true negative.

F-measure = average of recall and precision.
True positive = instances of spam messages in No. of correctly classified.
True Negative = instances of non spam messages in No. of correctly classified.
False Positive = Number of spam mail specify as non spam.
False negative = Number of non spam mail specify as spam[5].

Dataset
A collection of dataset is a spam and non spam messages. It consists 58 attributes and 4601 instances of spambase dataset. It is a benchmark dataset collected in UCI machine repository online portal [16-19]

Sample Database


```
@attribute word_freq_conference   numeric
@attribute char_freq_;            numeric
@attribute char_freq_(            numeric
@attribute char_freq_[            numeric
@attribute char_freq_!            numeric
@attribute char_freq_$            numeric
@attribute char_freq_#            numeric
@attribute capital_run_length_average numeric
@attribute capital_run_length_longest numeric
@attribute capital_run_length_total   numeric
@attribute spam                   {0,1}

@data

0,0.64,0.64,0,0.32,0,0,0,0,0,0,0.64,0,0,0,0.32,0,1.29,1.93,0,0.96,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.21,0.28,0.5,0,0.14,0.28,0.21,0.07,0,0.94,0.21,0.79,0.65,0.21,0.14,0.14,0.07,0.28,3.47,0,1.59,0,0.43,0
0,0.372,0.18,0.048,5.114,101,1028,1
0.06,0,0.71,0,1.23,0.19,0.19,0.12,0.64,0.25,0.38,0.45,0.12,0,1.75,0.06,0.06,1.03,1.36,0.32,0.51,0,1.16,
0,0.01,0.143,0,0.276,0.184,0.01,9.821,485,2259,1
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.85,0,0,1.85,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.92,0,0,0,0,0.64,0.96,1.28,0,0,0,0.96,0,0.32,3.85,0,0.64,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.88,0,0,1.88,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.15,0,0.46,0,0.61,0,0.3,0,0.92,0.76,0.76,0.92,0,0,0,0,0,0.15,1.23,3.53,2,0,0,0.15,0,0,0,0,0,0,0,0,0.15,
744,445,1257,1
```


5.Classification methods

J48 Classifier

J48 is a java algorithm and freely available then it is called open source. It is a decision tree at each node of tree the attribute again split the samples in subset. Represents a decision in every leaf node[4].

Naive Bayes classifier method

Naïve Baye's classifiers are used to give accurate results to find spam mail or not. It is classifier that uses the concept of conditional probability, and applying bayes theorem, it says that an evidence E and hypothesis H that Bears on that hypothesis [1].

$$P_r\left(\frac{H}{S}\right) - \frac{P_r\left(\frac{H}{S}\right) \cdot P_r(H)}{P_r(H)}$$

Where, $P_r[H]$ = probability of legitimate

$P_r[H/S]$ = probability of Ham on event spam [1].

Random Forest Tree

Random Forest Tree builds a decision tree that generates the best predictors. It generates a tree with forest of tree and every tree gives a specification in the forest over all the trees[4].

REPTree

Also a decision tree with capability to build fast in less time. It buids a decision tree and produces it by reduced error pruning. Only numeric attributes sort in this algorithm. Missing values are splitting the instances into pieces [20-22].

Random Tree

A random tree is a group of tree predictors, each predictor gives response with predictor values. In this algorithm, gives best result in spam classification with time and accuracy [4].

6.Conclusion and future work

Spam is a big problem in everyday life. So handling of spam mails is necessary. To specify the spam mails different methods are used to specifically identify the mail is spam or not that gives best solutions in terms of efficiency, less time and less error. In this paper Random tree, Naïve Baye's classifier, Random forest, and Reptree, J48 classifier are used. In Future, more number of different algorithms are used to identify the mail is spam or not to give perfect results with measures of accuracy and efficiency. Number of different software's are manufactures to classify the spam filtering approaches in future and classifiers are used like rapid miner, Matlab, Tangra etc to use different results and to check best results [1].

References

[1] Sharma AK, Sahni S. A comparative study of classification algorithms for spam email data analysis. International Journal on Computer Science and Engineering. 2011; 3(5):1890-5.

[2] Radhakrishnan A, Vaidhehi V. Email Classification Using Machine Learning Algorithms.

[3] Yitagesu ME, Tijare M. Email classification using classification method. International Journal of Engineering Trends and Technology (IJETT). 2016; 32(3):142-5.

[4] Deshmukh P, Rai Y, Kushwaha S. Identifying Malicious Behavior in MANET: A Survey. International Journal of Advanced Technology and Engineering Exploration. 2015; 2(4):43-8.

[5] Wu CT, Cheng KT, Zhu Q, Wu YL. Using visual features for anti-spam filtering. In Image Processing. ICIP 2005. IEEE International Conference on 2005 (pp. III-509). IEEE.

[6] Schneider KM. A comparison of event models for Naive Bayes anti-spam e-mail filtering. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics 2003 (pp. 307-14). Association for Computational Linguistics.

[7] https://en.wikipedia.org/wiki/Email_spam. Accessed 26 March 2018.

[8] https://archive.ics.uci.edu/ml/datasets/spambase. Accessed 26 March 2018.

[9] https://www.inf.ed.ac.uk/teaching/courses/iaml/lab/lab1.html. Accessed 26 March 2018.

[10] Wu CH. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Systems with Applications. 2009; 36(3):4321-30.

[11] Ma W, Tran D, Sharma D. A novel spam email detection system based on negative selection. In computer sciences and convergence information technology, 2009. ICCIT'09. Fourth international conference on 2009 (pp. 987-92). IEEE.
[12] Rekha SN. A Review on different spam detection approaches. International Journal of Engineering Trends and Technology (IJETT). 2014; 11(6):315-8.
[13] Nadaf SB, Gujar AD. A survey paper on spam mail detection using RFD. International Journal. 2016; 4(1):46-8.
[14] Agrawal N, Singh S. Origin (dynamic blacklisting) based spammer detection and spam mail filtering approach. In Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016 Third International Conference on 2016 (pp. 99-104). IEEE.
[15] Chae MK, Alsadoon A, Prasad PW, Sreedharan S. Spam filtering email classification (SFECM) using gain and graph mining algorithm. In anti-cyber crimes (ICACC), 2017 2nd International Conference on 2017 (pp. 217-22). IEEE.
[16] Easwaramoorthy S, Thamburasa S, Aravind K, Bhushan SB, Rajadurai H. Heterogeneous classifier model for E-mail spam classification using FSO feature selection method. In inventive computation technologies (ICICT), international conference on 2016 (pp. 1-6). IEEE.
[17] Harisinghaney A, Dixit A, Gupta S, Arora A. Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. In optimization, reliabilty, and information technology (ICROIT), 2014 International Conference on 2014 (pp. 153-5). IEEE.
[18] Kaur H, Sharma A. Improved email spam classification method using integrated particle swarm optimization and decision tree. In next generation computing technologies (NGCT), 2nd International Conference on 2016 (pp. 516-21). IEEE.
[19] Manjusha K, Kumar R. Spam mail classification using combined approach of bayesian and neural network. In computational intelligence and communication networks (CICN), 2010 international conference on 2010 (pp. 145-9). IEEE.
[20] Su CY, Shen DF, Lin GS. An image spam detection method. In consumer electronics-taiwan (ICCE-TW), 2017 IEEE International Conference on 2017 (pp. 71-2). IEEE.
[21] Tuteja SK, Bogiri N. Email Spam filtering using BPNN classification algorithm. In automatic control and dynamic optimization techniques (ICACDOT), International Conference on 2016 (pp. 915-19). IEEE.
[22] Vyas T, Prajapati P, Gadhwal S. A survey and evaluation of supervised machine learning techniques for spam e-mail filtering. In electrical, computer and communication technologies (ICECCT), IEEE International Conference on 2015 (pp. 1-7). IEEE.