Advances and Applications in Mathematical Sciences
Volume 20, Issue 12, October 2021, Pages 3087-3092

CLASSIFICATION OF E-MAIL SPAM WITH SUPERVISED MACHINE LEARNING - NAÏVE BAYESIAN CLASSIFICATION

J. PHANI PRASAD and T. VENKATESHAM

Information Technology
AGBS, Hyderabad, India
E-mail: jpprasad@hyd.amity.edu
tvenkatesham@hyd.amity.edu

Abstract

Spam in E-mail is the latest problem that every user is facing in the internet world, the email spam is just like an advertisement that is received by the client inbox without any prior intimation; to address this issue we have lot of spam filtering mechanisms which safe guards our mailbox. In this paper we are utilising the Naive Bayesian classifier for classifying spam mails. The data set used is Ling spam dataset for the classification of spam and non-spam E-mails.

Introduction

In the present scenario the unimportant E-mail messages are increasing in this internet which we name it as spam in general. The person sending those messages is a spammer. Spammers collect the data from e mails, websites and other social media platforms because of these spam e mails the computer resources and the bandwidth, CPU power of the email servers are effected much. Users who receive spam emails will feel very irritating also the financial loss is also very high for the people who are subjected to the scams in internet and other forms of fraud which is sent by unknown persons who send e mails like from reputed organizations with the intent of getting the personal information like credit card information or passwords, Bank cards Pin numbers. Latest statistics shows that spam messages accounted for

2010 Mathematics Subject Classification: 68Txx.
Keywords: E-mail, Spam; Classification; Naive Bayesian Classifier.
Received September 28, 2020; Accepted October 22, 2020

56.87% of E-mail traffic worldwide and the most familiar types of spam E-mails were healthcare and dating spam.

To efficiently handle the threat by E-mails the popular search engines like Google, Yahoo and Outlook has employed various types of Machine Learning Techniques like Neural Networks in its spam filters. The machine learning is capable of not only identifying the e-mails but also detect the spam part of E-mails using some pre-existing rules, with the help of the new rules what they have generated out of their own learning in the operation of spam filtering.

Statistics from Google revealed that between 50-70 percent of emails that Gmail receives are unsolicited mail. The different kinds of spam filtering techniques for spam filtering are given below:

- **Content Based Filtering Technique:** content-based filters evaluate words or phrases found in each individual message to determine whether an E-mail is spam or legitimate. The algorithms used are KNN, Random forest, Naive Bayes classification and Support Vector Machines.
- **Case Based Spam filtering Technique:** In this method both the spam and non-spam E-mails are identified and extracted, then pre-processing of the data and other steps are carried out and finally the machine learning model is used to separate the spam and ham mails.
- **Heuristic Based spam filtering Technique:** it is a rule based technique where it uses patterns to detect the E-mail to tell whether it is spam or not, various patterns which are
- Identical or same increases the score of the message, if any message score is less than a specified threshold then such email message is treated as spam.

E-mail Spam Filtering Architecture:

E mail has the following parts: one is header the other one is the body, the header part of e mail consist of addresses of sender and receiver whereas the body part consist of the actual information of the mail. Headers allow the

user to view the route the email passes through, and the time taken by each server to treat the mail. Before filtering of the data the information has to undergo some processing.

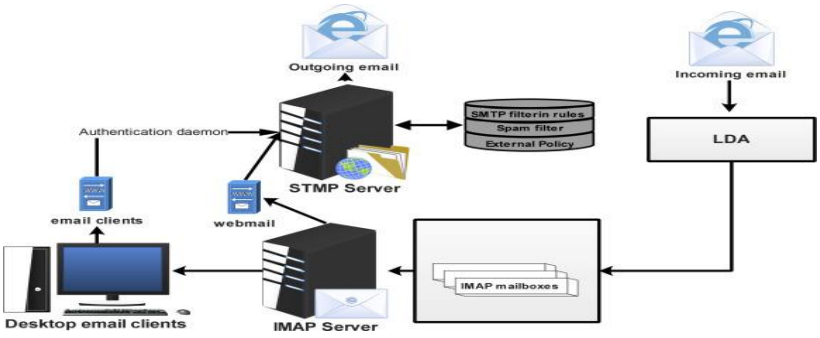
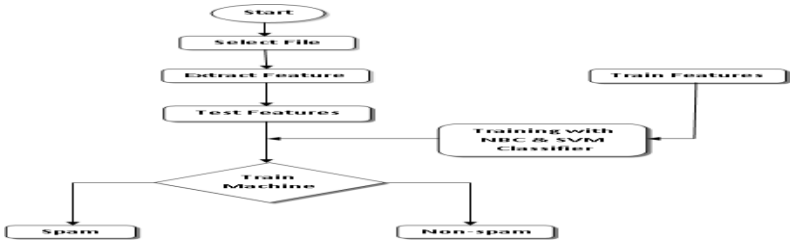


Fig.1. Email spam filtering Architecture.

The steps that are required in extraction of the data from an E mail message are as follows:

1. Pre-processing: when an incoming E-mail is received this particular step will be executed.
2. Tokenization: It is a process of dividing the words in to some serial symbols.
3. Feature Selection: When data size is large at that point of time the required features are extracted from large e mails or databases.



Navie Bayesian classification algorithm for e-mail spam classification

Step1. Input: E-mail Message Data set

Step 2. Parse each E-mail in to its component tokens

Step 3. compute probability to each token

$$S_p[W] = C_{spam}(W)/(C_{ham}(W)/(C_{ham}(W) + C_{spam}(W))$$

Step 4. save spam values in a database

Step 5. for each Message M do

Step 6. while (M not end) do

Step 7. scan message for next token Ti

Step 8. query database for spaminess

Step 9. compute the probabilities of message collected Sp[M] and Ha[M]

Step10. Compute the total message filtering signal: I[M]=f(Sa[M],Ha[M])

Step 11. $I[M] = I + Sa[M] - Ha[M]/2$

Step 12. if $I[M] >$ threshold then

Step 13. message is labelled as spam

Step 14. else

Step 15. message is labelled as non-spam

Step 16. end if

Step 17. end while

Step 18. end for

Step 19. return final email message classification (spam/not spam)

Step 20. end.

Description: In this method first we are taking the Data set and then we are pre-processing the data and then computing the probability of both spam and non-spam mails and storing that data in the database, This process is repeated for each and every token in the message M, then check the calculated component $I[M]$ with a common threshold which is already set and if $I[M]$ is greater than threshold then message is treated as spam otherwise ham.

Results

Here in our study we are taking the Data set as Ling spam data set in which we had taken totally 960 e mails out of which 700 are used for training data set and 260 for test data set. Again in 700 we had considered 350 mails as spam and 350 as non-spam and in 260 we considered 130 as spam and other 130 as non-spam mails.

Here is the result of four split ted and trained Data set and the results are compared between the SVM classifier and the Naïve Bayes classifier

Train Dataset	Accuracy of SVM	Accuracy of NBC	Error rate of SVM	Error rate of NBC
Dataset-50	67	7	0.25769	0.026923
Dataset-100	24	6	0.092308	0.023077
Dataset-400	9	6	0.034615	0.023077
Dataset-700	6	5	0.023077	0.019231

Conclusion

E-mail spam is a major concern now a days in the usage of internet and search engine type of communication, detecting and filtering of spam is the need of the hour by using many machine learning algorithms like supervised and unsupervised we can solve the E-mail spam issue and it can be extended to Deep learning and Neural Networks in future.

References

[1] Prabin Kumar Panighrahi, A Comparative Study of Supervised Machine Learning Techniques for Spam E-Mail Filtering, Fourth International Conference on Computational Intelligence and Communication Networks 2012.

[2] Dada, Bassi and shafi, Machine Learning for E-mail spam filtering: review, approaches and open research problems, science direct, 2019.

[3] Priyanka sao and k. Prashanthi, E-mail spam classification using Naïve Bayesian Classifier, IJARCET l4(6) 2015.

[4] J. R. Mendez and F. Diaz, A comparative performance study of feature selection methods for the anti-spam filtering domain, advances in data mining (2007).

[5] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras and C.D. Spyropoulos, An evaluation of naive bayesian anti-spam filtering, Proceedings of 11th European Conference on Machine Learning (ECML 2000), Barcelona (2000), 9-17

[6] S. Abu-nimeh, D. Nappa, X. Wang and S. Nair, A comparison of machine learning techniques for phishing detection, Proceedings of the Anti phishing Working Groups 2nd Annual eCrime Researchers Summit, New York, USA (2007).

[7] I.Androutsopoulos, J. Koutsias, K.V. Chandrinos and C.D. Spyropoulos, An experimental comparison of naïve Bayesian and keyword-based anti-spam filtering with personal e-mail messages, Proc of the Ann Int. ACM SIGIR Conf on Res and Devel in Inform Retrieval (2000).

[8] K.P. Clark, A Survey of Content-Based Spam Classifiers, 2008.

[9] T. Saravanan, A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm, 2010.

[10] I. Biju, wendy, Implementing spam detection using Bayesian and porter stemmer keyword stripping approaches, TENCON 2009-2009 IEEE Region 10 Conference (2009).