Spam Filtration using Boyer Moore Algorithm and Naïve Method

Aastha Baranwal
Student, Department of CSE
Galgotias College of Engineering and Technology, Greater Noida

Gunjan Gaur
Student, Department of CSE
Galgotias College of Engineering and Technology, Greater Noida

Akanksha Bhasker
Student, Department of CSE
Galgotias College of Engineering and Technology, Greater Noida

Rishabh Jain
Assistant Professor, Department of CSE
Galgotias College of Engineering and Technology, Greater Noida

ABSTRACT
Emails are primarily being used for transporting information in a quicker and well-organized way. They are favored in professional as well as personal space because of its attribute of time saving and trading data over huge distances. According to the statistics of the past few years, forgery and fraudulent activities are frequent in emails and these are categorized as 'spams'. Spam emails utilize time, transmission capacity and storage section; hence it is essential to spot these mails to shield our treasured data and time from being distorted. There are numerous approaches that have been formulated to screen the emails and organize them as spam and non-spam.

The motive of scripting this paper is to analyze recent works done in spam detection and also present a new technique using graylist filter, Boyer Moore string searching algorithm and Naïve Bayes algorithm. Also, observe its working in contrast with traditional Naïve Bayes algorithm.

General Terms
Spam Filtration, Graylist Filter, Naïve Bayes, Boyer Moore Algorithm, String Matching.

Keywords
Spam, Ham, Tokenization, Classifier.

1. INTRODUCTION
Email is a method of sharing knowledgeable data across users using electronic devices. Emails are put into service in every field to transfer the vital information, but many a times users receive emails which are useless to them. Some organizations use emails as a channel to promote their services and products, or to extricate the privileged information from the user. Covertness of data can be compromised with the use of spams and fraud emails. Spam emails take up transmission capacity, time and storage area and fraud emails try to pull out sensitive information from the user by unjust means. There are many spam filters (like, ML based, Check-sum based and memory-based) that are employed to confront the problem of spams. Since new means of evasion are often introduced, new procedures to diagnose spam and fraud emails need to be discovered.

The email classification into spam and non-spam can be done using non-machine learning techniques and machine learning techniques. The white-list/black-list, graylist, signatures, email header analysis are some of the non-machine learning techniques used in classification of emails [1]. Machine based techniques are also known as content-based techniques and have higher performance as compared to that of the non-machine based. Currently, there are many machine-based classifiers so far in use such as classifiers based on Bayesian Algorithm, SVM (Support Vector Machine), MLP (Multi Layer Perceptron) approach, K-Nearest Neighbor, Logistic Regression, k-Means etc. Also, email can be assigned to a class (either spam or ham) by string matching algorithms like Knuth-Morris-Pratt, Boyer Moore, Rabin-Karp, etc.

Many classifiers are set up using single algorithm and give decent performance. However, the integration of machine based and non-machine based techniques can also be used. Higher performance can be achieved by merging two or more classifiers.

In this paper, we are surveying recent proposed works in this field and inspired from all above, we are introducing a classifier model using Naïve Bayes classifier along with Boyer Moore algorithm to classify emails into spam and ham. Also an overview of the functionality of the application is given.

This paper is divided into nine sections. In the second part of this paper, we have mentioned the literature survey that is possibly done in this field concerning our topic. The proposed work and related algorithms are mentioned in third and fourth sections of this paper respectively. In the fifth section the methodology of implementation of the idea is given. In the sixth section the result of the research is described. And finally we have concluded our research in seventh section.

2. LITERATURE SURVEY
In 2017, Akash Iyengar, G. Kalpana, Kalyankumar.S, S.GunaNandhini emphasized in integrated approach of detecting spam for multilingual mails. They used Bayesian classifier and graylist filter on Gmail and Yahoo dataset and accuracy increased by 1% over traditional approach i.e., 97.3%. Their work was inspired from a research conducted by Sunil B. Rathod et al. [2]

V. Suganya in the year 2016 published a review paper on "A Review on Phishing Attacks and Various Anti Phishing Techniques" in International Journal of Computer Applications. In that paper, she talked about some existing work on anti-phishing. [3]

In 2015, Sunil B. Rathod and Tareek M. Pattewar emphasized Bayesian approach for classifying spam and legitimate mails using supervised learning across features extracted applying the Bayesian classifier, they experimentally demonstrated that spam mails can be detected with an accuracy of more than 96.46% with respect to real world Gmail datasets. [4]

In 2015, Lin Li and Chi Li studied the current status of spam filtering technology, and selected Naive Bayes algorithm as spam filtering algorithm. In this paper, they did a detailed study and analysed Naïve Bayes classifier algorithm, focusing on the feature selection methods TF-IDF, they carried out research, and proposed improvements to enhance high-frequency words category weights, designed to improve the TF-IDF method based on Naive Bayes spam filter. [5]

R. Malarvizhi and K. Saraswathi in the year 2013 published a paper on "Content-Based Spam Filtering and Detection Algorithms- An Efficient Analysis & Comparison" in International Journal of Engineering Trends and Technology. In this paper they compared some classifiers like Adaboost classifier, Fisher-Robinson Inverse Chi-Square Function, Naive Bayes classifier and k-nearest neighbour for spam filtering. They reached to a conclusion that Naive Bayesian classifier is the efficient technique among the discussed techniques to create a spam filter. [6]

Omar Saad, Ashraf Darwish, and Ramadan Faraj talked about how to determine whether an email is a spam. They also gave a brief overview of the underlying theory and implementations of the algorithms like Naive Bayesian classifier, Support Vector Machine, Artificial Neural Network, k-nearest neighbour classifier, and Artificial Immune System classifier. [7]

In 2010, Zhengda Xiong gave a composite Boyer Moore algorithm for string matching problem. They gave analysis to several classical algorithms, KMP, BM and their improvements. Then, by compositing the main method of the BM algorithm, they proposed a new algorithm — the Composite BM algorithm (CBM). The result showed the efficiency of CBM is higher than of BM. [8]

In 2017, Kajaree Das and Rabi Narayan Behera explained the concept and evolution of Machine Learning, some of the popular Machine Learning algorithms and tried to compare most popular algorithms based on some basic notions. This journal gave us a basic understanding of the domain and its significance in solving the real world problems. [9]

3. PROPOSED WORK

In this paper we are employing a classifier for spam filtration which is using Naïve Bayes method to classify any new mail into spam or non-spam based on previous mails and Boyer Moore Algorithm to match the text with pattern and further improves the performance of the classifier by showing the percentage of spam content, entropy and information gain in emails.

4. RELATED ALGORITHMS

4.1 Naïve Bayes Classifier Algorithm

The Naïve Bayes Classifier works on Bayes theorem and it assumes that all the future vectors in context are mutually independent. For the classification of a new email into spam or ham, the stored statistical data helps to determine the probability of the email which will assign the relevant category to that email.

Table 1: Comparison between Gaussian NB, SVM and Decision Tree [9]

Algorithm	Training Time	Prediction Time	Accuracy
Naive Bayes (Guassian)	2.708	0.328	0.692
SVM	6.485	2.054	0.6565
Decision Tree	454.069	0.063	0.69

An efficient algorithm is chosen based on the dataset and the domain to which it is applied.

4.2 Boyer Moore Algorithm

Table 2: Comparison table for various string searching algorithm [10]

Algorithms	Search Type	Space	Time Complexity	Approach
Brute-Force	Prefix	none	O((n-m+1)m)	linear searching
Rabin-Karp	Prefix	O(1)	Θ(m),Θ(n+m)	hashing based
Boyer-Moore	Suffix	Θ(m)	O(m+|Σ|),O(n)	heuristic based
Knuth-Morris-Pratt	Prefix	Θ(k)	O(m),O(n+m)	heuristic based

As concluded from above table, Boyer Moore algorithm is better string searching algorithm.

In this algorithm, we match a pattern with the text string but in the backward direction, unlikely traditional string matching algorithm.

Step 1: The pattern is preprocessed by making a bad character shift table which tells about the number of character shifts in case of a mismatch.

Step 2: Pattern is compared with a substring in the text of the same length as that of pattern but from the rightmost direction.

Step 3: According to the table we shift the pattern across the text string and thus find the pattern in the string if it exists.

4.3 Graylist Filter

Generally, many spammers attempt to send a group of junk mail only once. In this filter, the receiving server rejects mails from unknown users the first time and sends a message of failure to the sender. If the sender again makes an attempt to send mail (done by legitimate senders mostly), then the filter presumes that the mail is not spam and lets it propagate to the recipient and then the filter adds the email id of both sender and recipient to the allowed users list.

5. METHODOLOGY

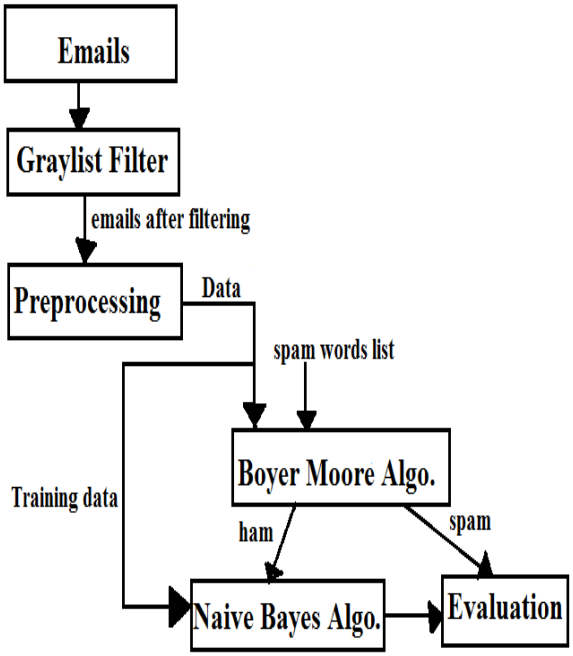

Figure 1: Flowchart of Implementation

5.1 Preprocessing of Emails

Preprocessing of emails includes various steps:

1. Words with length<=2 are removed like- is, at, to, if, an, as, do, etc. However, words having 3 letters are also useless like was, for, the, etc.
2. Alphanumeric words are removed as they do not generally repeat in the email.
3. There are certain words which do not add meaning to the spam detection, called stop words like then, there, can, by, etc. and are removed.
4. Verbs are altered into their first form and plural nouns are improved to their singular form. This is called Stemming.

After pre-processing, the email is given as an input to Boyer Moore algorithm to match the text with spam words already defined. This step classifies a mail into spam or non-spam based on the threshold specified in the programming module using Boyer Moore algorithm. After this step, spam mails are sent for evaluation and ham mails are further sent for classification using Naïve Bayes classifier. Ham emails are sent for processing as the content of the ham mail did not match with the stored dictionary of the spam words. Hence it is classified as ham email. To check for unprecedented words content in the mail we require this step.

Naïve Bayes classifier learns from the training data provided to it and classifies the email as spam or ham.

After the classification, percentage of spam and ham is calculated thereby calculating the entropy or information gain of the mail.

The logic that classifies the mail into spam or normal is shown below-

double probofspam = matches / tot;

double notmatched = tot - matches;

double probofham = notmatched /tot;

here, the 'matches' shows the number of the word matched in the list of spam words and 'tot' shows total number of words in text mail.

double entropyofspam = Math.Log(probofspam, 2);

double entropyofham = Math.Log(probofham, 2);

entropy of the spam content and normal content in the mail is calculated and displayed each time the mail is sent.

6. RESULT AND ANALYSIS

The aim of the research to filter out the spam emails is achieved and the model shows the probability of the spam content in an email and also calculates the entropy.

Each time the mail is sent, sender's id is compared with the spam id list created by the admin as suggested by users. If the mail is sent from the user listed as spammer, it is blocked. In the other case where sender id is a normal id then the mail is processed for the spam content in it using both the algorithms. When the mail is sent, probability of spam and ham, and also entropy of spam and ham is shown along with sent popup message box.

The increase in the performance has been observed as more layers of classification are involved.

Table 3: This comparison table shows that our method of classifying mails is better than traditional Naïve Bayes

Algorithm	Training Time	Prediction Time	Accuracy
Naive Bayes	same	same	less accurate
Naive Bayes (with Boyer Moore)	same	same	more accurate

7. CONCLUSION AND FUTURE SCOPE

In this research we have extended the classical Naïve Bayes classifier with a non-machine learning technique and a pattern matching algorithm. After analyzing the performance of the proposed model we can conclude that the Boyer Moore algorithm enhances the performance in the form of accuracy of the classification process.

We have used this classification model only for classifying text emails, further addition of pictures, video and audio for classification. For pre-processing, stemming can also be used. We can employ the advanced version of the Boyer Moore algorithm, which is Composite Boyer Moore algorithm and check its function in combination with Naïve Bayes algorithm.

8. ACKNOWLEDGEMENT

We would like to thank our teachers and colleagues for their valuable comments and suggestions that greatly improved the paper. We are thankful to our guide Mr. Rishabh Jain for his motivation and valuable advice. We are also immensely grateful to our HOD and coordinators for their support. Although errors present in this paper, if any, are our own and

should not deteriorate the reputation of these esteemed persons.

9. REFERENCES

[1] S. Aski and N. K. Sourati, -Proposed efficient algorithm to filter spam using machine learning techniques, in International Journal of Innovative Research in Computer and Communication Engineering, 2017

[2] Akash Iyengar, G.Kalpana, Kalyankumar.S, S.GunaNandhini, "Integrated Spam Detection for Pacific Science Review- A Natural Science Engineering- Elsevier., vol. 18, no. 2, pp. 145-149, 2016

[3] V. Suganya, "A Review on Phishing Attacks and Various Anti Phishing Techniques" in International Journal of Computer Applications, 2016

[4] Sunil B. Rathod, Tareek M. Pattewar, "Content Based Spam Detection in Email using Bayesian Classifier" in International Conference on Cryptography, Security and Privacy, 2015

[5] Lin Li, Chi Li, "Research and Improvement of a Spam Filter based on Naïve Bayes" in International Conference on Intelligent Human-Machine Systems and Cybernetics, 2015

[6] R. Malarvizhi, K. Saraswathi, "Content-Based Spam Filtering and Detection Algorithm- An Efficient Analysis & Comparison" in International Journal of Engineering Trends and Technology, 2013

[7] Omar Saad, Ashraf Darwish, Ramadan Faraj, "A survey of machine learning techniques for Spam filtering" in International Journal of Computer Science and Network Security, 2012

[8] Zhengda Xiong, "A Composite Boyer-Moore Algorithm for the String Matching Problem" in International Conference on Parallel and Distributed Computing, Applications and Technologies, 2010

[9] Kajaree Das, Rabi Narayan Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications." Multilingual Emails" in International Conference on Information, Communication, and Embedded Systems, 2017

[10] Akhtar Rasool, Amrita Tiwari, Gunjan Singla, Nilay Khare, "String Matching Methodologies: A Comaparative Analysis" in International Journal of Computer Science and Information Technologies, 2012

[11] Anjali Sharma, Manisha, Dr. Rekha Jain, Dr. Manisha, "Data Preprocessing in Spam Detection" in International Journal of Science Technology and Engineering, 2015

[12] William S. Yerazunis, Shalendra Chhabra, Christian Siefkes, Fidelis Assis, Dimitrios, "A Unified Model of Spam Filtration" in Mitsubishi Electric Research Laboratories, 2005

[13] Saadat Nazirova, "Survey on Spam Filtering Techniques" in Communications and Network, Scientific Research, 2011

[14] Mamoru Kato, Joseph Langeway, Yimin Wu, William S. Yerazunis, "Three Non-Bayesian Methods of Spam Filtration" in TREC, 2007

[15] Neha Roy, Rishabh Jain, "Virtual Machine Scheduling on Clouds Using DVFS" in International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277 128X, Volume 5, Issue 5, May 2015.

[16] Pooja Ahlawat, Poonam, Rishabh Jain, "An Improvement to Life of Wireless Sensor Network Using Leach Design a Cluster Head"- IJCSMS (International Journal of Computer Science & Management Studies) ISSN(Online): 2231-5268, Volume 15, Issue 06, June

[17]