A Survey on Spam Filtration and Fraud Detection in Emails using Machine Learning Algorithms

[1]Aastha Baranwal, [2]Akanksha Bhasker, [3]Gunjan Gaur, [4]Rishabh Jain

[1]Student, [2]Student, [3]Student, [4]Assistant Professor
[1]Department of CSE,
[1]Galgotias College of Engineering and Technology, Greater Noida

Abstract : Emails are being widely used for communicating information in an efficient and faster way. As they save up our time and helps in exchanging information over large distances, they are preferred in professional and personal space. But over the past few years, there have been instances reported of forgery, fraudulent activities and these emails are categorized as 'spams'. Spam emails take up time, bandwidth and storage area, so it is very important to detect these emails to prevent our valuable information and time from being misused. There are various techniques that have been devised to filter the emails and classify them as spam or not.

The purpose of writing this paper is to review recent works done in spam and fraud detection and filtration in emails. Also, a new model using grey list filter, K-means clustering algorithm along with Random Forest Classification Algorithm is proposed to do the same.

***IndexTerms*- Spam filtration, Greylist Filter, K-means, Random Forest Algorithm, Classification.**

I. Introduction

Email is a method of sharing information among users with the help of electronic devices. Emails are employed in every field to communicate the important information, but many a times user receives emails which is of no use to him. Some organizations use emails as a medium to promote their services and products, or to extract the confidential information from the user. Secrecy of data can be compromised with the use of spams and fraud emails(phishing).Spam emails take up bandwidth, time and storage area and fraud emails try to extract sensitive information from the user by unfair means. There are many spam filters(like Bayesian filters, Checksum-based filters, machine learning based filters and memory-based filters) that are employed to encounter the problem of spams but they have high false positives. Since new methods of evasion are introduced frequently, new ways of detecting spam and fraud emails need to be discovered.

The email classification into spam and ham can be done using non-machine learning techniques and machine learning techniques. The white-list/black-list, grey-list, signatures, email header analysis are some of the non-machine learning techniques used in classification of emails [1]. Content-based techniques are machine-based and have higher performance as compared to that of the non-machine based. Presently, there are many machine-based classifiers already in use such as classifiers based on Bayesian Algorithm, SVM (Support Vector Machine), MLP (Multi Layer Perceptron) approach, K-Nearest Neighbour, Logistic Regression, k-Means etc.

Many classifiers are built using single algorithm and give good performance. However, the integration of machine based and non-machine based techniques can also be used. Higher performance can be achievedby integrating classification with clustering techniques.

In this paper, we are surveying recent proposed works in this field and inspired from all above, we are proposing an idea for integrating grey list technique, k-means clustering and random forest algorithm to classify spam and ham.

This paper is divided into five sections. In the second part of this paper, we have mentioned all the recent work that is done in the field related to this research paper. The proposed work and related algorithms are mentioned in third and fourth sections of this paper respectively. In the fifth section the methodology of future implementation of the idea is given.

II. Recent Work

In 2017, Dr.Priti and Uma gave the performance comparison analysis of various pre-existing email classification technique along with their limitations.[2]

In 2017, Kajaree Das and Rabi Narayan Behera surveyed the recent machine learning concepts, algorithms and applications. They compared the performance of Naïve Bayes, SVM and Decision Tree based on some basic notion i.e., training time, prediction time and accuracy. [3]

In 2015, Sunil B. Rathod and Tareek M. Pattewar used Bayesian classifier to classify emails into ham and spam. They did the classification on the basis of content of the body of email. The testing dataset was derived from Gmail consisting of spam and legitimate mails. For preprocessing, HTML tag removal, Stopword Removal, Tokenization and word frequency. On application of Bayesian classifier, they experimentally demonstrated that spam emails were detected with an accuracy of more than 96.46% with respect to real world gmail datasets. [4]

Inspired from above work, in 2017, Akash Iyengar, G. Kalpana, Kalyankumar S., S. Guna Nandhini emphasized on integrated approach of detecting spam for multilingual mails. They used bayesian classifier and grey list filter on gmail and yahoo dataset and accuracy increased by 1% over traditional approach i.e. 97.3%. [5]

In 2017, P. Priyatharsini, Dr. C. Chandrasekar compared the performance of six different decision trees on UCI email dataset using WEKA mining tools. Preprocessing of the dataset was done due to presence of noise and missing values. Feature reduction was done using techniques Relief F, Chi Square attribute level. Performance of different algorithms were shown before and after reduction.LMT and RF gave best results but random forest comparatively had the best performance with 99% accuracy, zero false positive rate and least error rate. [6]

In 2017, Harjot Kaur and Er. Prince Verma refined a supervised approach, MLP using a fast and efficient unsupervised approach, K-Means for the detection of spam emails by selecting best features using N-Gram technique. K-means approach was used to cluster the data set into ham and spam and MLP was used as a classifier. Simple MLP and SVM which were initially carried out showed the accuracy of78.09% and 64.66% respectively, evidently MLP shows higher performance than SVM. N-Gram technique was used to enhance the performance of MLP. Performance of the MLP was boosted to 97.53% for Bi-Gram, 98.23% for Tri-Gram and 99.00% for 4-Gram. N-Gram helped in choosing the best features from the large dataset. Comparative analysis of performance showed that this proposed model gave better results over simple MLP, simple SVM, and N-gram based K-SVM. [7]

In 2016, Manish Kumar conducted an experiment for spam mail filtering task on the dataset obtained from UCI Machine Learning Repository separately using ten machine learning algorithms with ten-fold cross validation. The result showed that classifier Random Forest is better amongst every other classifier (including SVM, which was reported as the better performing classifier by the previous studies) with AUC (area under ROC curve), accuracy and MCC(Mathews Correlation Coefficient) value up to 0.987,0.955 and 0.906 respectively. [8]

In 2015, Zhiqiang Ma, Rui Yan, Donghong Yuan and Limin Liu designed a classification model for distinguishing spam mails from ham mails. Using K-means algorithm to cluster the mails, they extracted a typical example of spam mail. K-means algorithm was used to convert the unbalanced data into the balanced. Then SVM model is used to classify the emails into ham and spam mails. This is referred as K-SVM classification model which is significantly better than SVM model. This is designed to overcome the problem of high false positive rate and time consumption in treatment of the data. They achieved high classified efficiency and generalization performance as compared to SVM model. [9]

III. PROPOSED WORK

In this paper we have proposed a hybrid approach to classify the mails into spam and ham emails. An integrated spam filter consisting of Grey list filter, k-Means clustering and random forest algorithm will be used as these individually produce better results. So, their integration might give better results with least false positives.

Greylist filtering is an initial step to deal with an email and in this it will also be decided whether further processing is required or not. Email will be classified as spam or ham with the help of classifier based on random forest algorithm. The classifier will be trained based on the features of a typical spam email obtained by clustering emails using k-means algorithm.

IV. RELATED ALGORITHMS

This section gives a brief idea on the algorithms that we will be using in the process of bringing this proposed work into reality.

3.1 Greylist 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 or IP address of both the sender and recipient to the allowed users list.

3.2 k-Means Clustering

Initially considering a typical spam and ham email as centers clustering of dataset is done into two clusters- ham email and spam email. The centre keeps on changing when new data is added to any of the cluster each time. This unsupervised machine learning method is used to label any new data. In this technique, Euclidean distance between new data and centers is calculated and accordingly assigned to one of the clusters. Then a new centre for that particular cluster is calculated by taking means of the data present in that cluster.

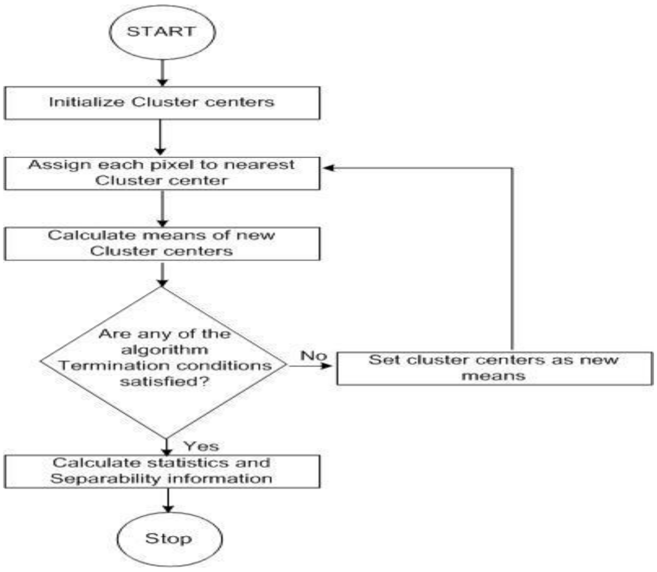
START
Initialize Cluster centers
Assign each pixel to nearest Cluster center
Calculate means of new Cluster centers
Are any of the algorithm Termination conditions satisfied?
No
Set cluster centers as new means
Yes
Calculate statistics and Separability information
Stop

Figure 1 k-means clustering algorithm flowchart [10]

3.3 Random Forest

This algorithm creates a bunch of decision trees as is the name suggests 'forest'. It is said to be random as it trains each decision tree with different subset of training data. Each node of each decision tree is split using a randomly selected feature from the dataset. Through this, the algorithm is able to create models that are not related to each other. Thus, this diversification helps to classify test data keeping in mind different features. At last majority voting decision strategy is done for classification of test data.

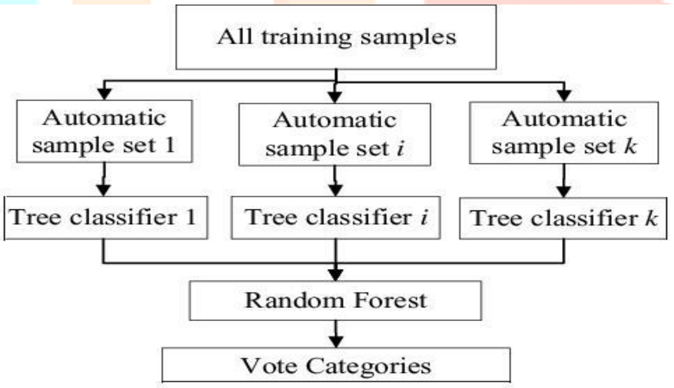

Figure 2 random forest algorithm flowchart [11]

V. METHODOLOGY

We have drawn an outline of the implementation structure of the model that will be employed to classify the emails as spam or ham. Firstly the sample emails will be collected from Gmail, yahoo accounts. Clustering will be performed to get a typical example of spam email and ham email. Then the feature extraction will be performed to select a number of features for making a decision about the type of the mail. Then Grey List filter will be used to filter out the emails received from the Senders labeled as fraudsters or spammers. This helps in the optimization of the whole process as it removes tedium by filtering out the obvious spam emails

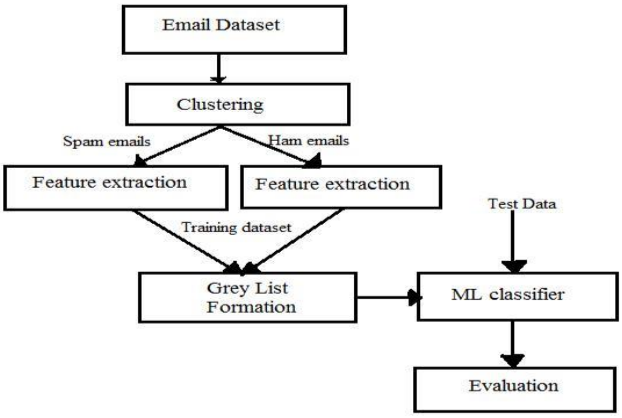

Figure 3 flowchart of implementation

The ML classifier based on the Random Forest algorithm will receive both the Test data and the training data set as this will help it in learning from its past experience. The output from the classifier will be evaluated on performance criteria and the spam will be categorized as spam email or ham email.

A comparative analysis will be done with the existing classifiers on the performance decisive factors such as time, accuracy, false positive rate etc., while evaluating the performance of the classifier.

VI. Acknowledgment

We would like to thank our teachers and colleagues for their valuable comments and suggestions that greatly improved the paper. 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.

References

[1] A. S. Aski and N. K. Sourati, ―Proposed efficient algorithm to filter spam using machine learning techniques, Pacific Science Review- A Natural Science Engineering- Elsevier., vol. 18, no. 2, pp. 145–149, 2016.

[2] Dr. Priti and Uma, "Performance Analysis, Comparative Survey of Various Classification Techniques in Spam Mail Filtering",Oriental Journal of Computer Science and Technology, Vol.10, No.3, 2017.

[3] Kajaree Das, Rabi Narayan Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017.

[4] Sunil B. Rathod, Tareek M. Pattewar, "Content Based Spam detection in Email using Bayesian Classifier", ICCSP Conference, 2015.

[5] Akash Iyengar, G. Kalpana, Kalyankumar.S, S.Guna Nandhini, "Integrated Spam Detection for Multilingual Emails", International Conference on Information, Communication and Embedded Systems, 2017.

[6] P.Priyartharsini, Dr. C.Chandrasekar, "Email Spam Filtering in Data Mining", International Journal of Engineering Science and Computing, Vol. 7, No. 11, November 2017.

[7] Harjot Kaur and Er. Prince Verma, "K-MLP Based Classifier for Discernment of Gratuitous Mails using N-Gram Filtration", International Journal of Compter Network and Information Security, 2017

[8] Manish Kumar, "International Journal of Innovative Research in Computer and Communication Engineering", Vol.4, Issue 3, March 2016.

[9] Zhinqiang Ma, Rui Yan, Donghong Yuan and Limin Liu, "An Imbalanced Spam Mail Filtering Method", International Journal of Multimedia and Ubiquitous Engineering, Vol. 10, No.3, 2015.

[10] https://nptel.ac.in/courses/105104100/lectureD_28/images/13.gif

[11] https://www.researchgate.net/figure/Random-Forest classification-principle_fig2_310467682

[12] Poonam, R Jain, "Review Paper for Improvement Life of Wireless Sensor Network using Leach Design", International Journal of Computer Science and Management Studies, Vol. 15, Issue 5, May 2015.

[13] S Gaba, P Ahlawat, R Jain, "Implementation on Fuzzy Approach to Query Traditional Database", International Journal of Computer Science and Management Studies, Vol. 15, Issue 6, June 2015.

[14] Poonam, R Jain, "An Improvement to Life of Wireless Sensor Network using Leach Design a Cluster Head", International Journal of Computer Science and Management Studies, Vol. 15, Issue 6, 2015.