DETECTION OF SPAM EMAILS USING HYBRID ALGORITHMS

[1]V.Sowmya Sri, [2]K.Punith, [3]Praveen Tumuluru

ABSTRACT--*As E-mail is the fastest mode of sending information all over the world, many people are trying their best to misuse the services. This misuse includes sending of fake emails also known as spam mails. It is either not easy to detect what is spam email and delete the emails because its hard to find out spam emails among them. Many people without having knowledge about what is an spam email when they come across such mails they either delete or remain same in their mail list. Mainly due to marketing and advertising agents emails send by them are* ***irregular*** *and contains unwanted information. These mails cause more storage in mails and time waste. There are some mails which can hack our system by clicking any links provided by them. By this many of the email accounts get hacked and cause great loss to the particular person. E- mail has established a major place in users life. Mails square measure used a significant and vital mode of communication in each personal and skilled aspects of ones life. The fast increase within the variety of account holders over the past number of decades conjointly the increase in mail volume have also made some serious issues. The communications square measure called ham and spam emails. Spam emails square measure is spreading at a huge place from the past few decades. Such spam emails square measure unlawful and unwanted emails that will contain virus, malicious codes, ads or threats. Machine Learning wont allow to screen the spam email mechanically at awfully sensible peace current days.This major problem has created a necessity for reliable and economical anti-spam filters that split the email into spam or ham messages. Spam filters keep the user from delivery spam emails into the inbox. Email spam filters can filter emails either on content or header base. Specific spam filters square measure is classified into 2 teams, particularly machine learning and non-machine learning. In this paper hybrid algorithms are used to detect spam emails. Machine Learning algorithms such as AdaBoost classifier, Gradient Boosting Classifier, Count Vectorizer and Naive Bayes Classifier are some of the algorithms used in prediction of spam emails. . By calculating the accuracy of each and every algorithm on given information the algorithms with high accuracy are combined for further process. A mail id with password is provided and lists of mails are also provided for detecting spam messages. After detecting it provides output with a graph providing the number of ham and number of spam messages detected.*

Keywords--*E-mail Spam, Classification, Spam Filter.*

I. INTRODUCTION

E-mail is the fastest mode of information transfer. It is because email is less cost and more efficient. This is misled by spam mails. These are the mails send by some users which contains unwanted information. These

[1] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Andhra Pradesh)

[2] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Andhra Pradesh)

[3] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Andhra Pradesh)

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

users are called as spammers. These mails are send one after the other even the same mail is forwarded many times. This kind of mails may irritate the user and make them think that email is irregular and they have a chance of quitting using mail.

In this paper several techniques to detect spam mails and they are classified as ham and spam. This process of spam detection mainly focuses on email address, content of the mail and subject of the mail. Spam mails are detected as in the way that they are different from remaining emails and they are since considered as spam emails. Knowledge engineering deals with IP address, it showed best results. Next technique is machine learning algorithms. The process of detection is an email id with password is provided and data is pre-processed. Pre-processed includes removing stop words(the words when removed remains the same meaning of the sentence and doesn’t show more change of meaning), work tokenization(the sentence is tokenized as tokens). The pre-processing stage reduces the dimension of details, and then extracts features in the format of a word bag. Recently huge email, also called as spam, is becoming a big internet issue. Spam is considered as waste of time and energy, space for storage, and bandwidth for communication.

The spam email problem has been on the rapid increase for years. At the moment, advanced e- mail filtering appears to be the most powerful way to combat spam and there is a close struggle between spammers and spam-filtering systems. Most of these spam can be controlled accurately about few years earlier by preventing e-mails originating from certain addresses or shutting out messages from those subject lines. Spammers began using many tricky approaches to overcome filtering techniques, such as using random sender addresses and/or applying random characters to the start or end of the subject line of the message. The two basic strategies used in e-mail filtering are information technologies and machine learning.

In software engineering approaches a set of rules should be defined according to the pattern of how emails are classified as spam or ham. A collection of these rules should be generated either by the filter user or by any other authority. No promising results are achieved by applying this process as the rules must be frequently modified and maintained, which is a waste of time and is not easy for most users. Machine learning approaches are more reliable than methodology to software engineering it does not allow any guidelines to be set down.

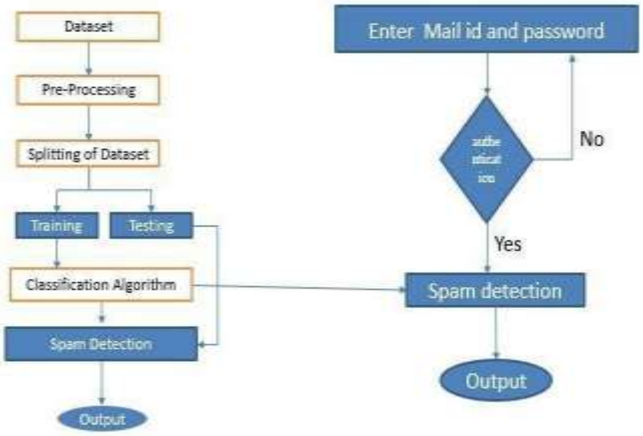

Figure 1: Flow diagram

These tests are then a series of teaching tests, a collection of pre- classified e-mail addresses. Afterwards, a complex algorithm is used to learn from these email messages the classification rules. Machine learning approach has been studied extensively and there are many algorithms that can be used in filtering e-mails. A significant work on anti-spam tactics was undertaken to tackle this problem, and different forms of anti-spam applications were developed and used by email users. The techniques used in spam filters include both manual and automated approaches. Negative lists of spammers, list of authenticated senders and selected list of terms in email text or subject are found in manual methods for the creation of anti spam filters. Machine learning methodology, a superior technology relative to manual approaches, has been implemented in recent years to automatically recognize and identify spam emails.

The algorithms learn from the training dataset in supervised or inductive machine learning, which comprises all inputs and outputs and a model is generated. The pattern is then checked for classifying new samples. The success is in two groups in the case of binary classification.

E-mail spam filtering is one of the most important study areas of recent days. Users may connect whole domains or email addresses, or usable domains. An interesting alternative is an automated white list management tool which removes the need for users to individually enter accepted addresses on the white list and ensures that mail from particular senders or domains is never marked as spam. Almost all of the techniques for spam filtering are focused on approaches for text categorization. So spam filtering transforms to a question of sorting. Rules for extracting function vector from emails are presented in our research. Provided that the features of prejudice are not clearly known, the use of machine learning strategies is more convenient. Machine Learning algorithms such as AdaBoost, Gradient boosting, Count Vectorizer, Extra Tress Classifier are some of the algorithms. By calculating their accuracy on the application the algorithms with high accuracy are considered further for hybrid algorithms. They are finally classified as ham, spam and displayed in the form of a graph.

II. METHODOLOGIES

2.1 NAÏVE BAYES CLASSIFIER

The basic yet powerful classifier Naïve Bayes classifier used in various knowledge processing applications including Natural Language processing, Knowledge extraction respectively. A Bayesian theorem are the theorems based on Naïve Bayes and is appropriate when the inputs are high in dimensionally. Naïve Bayes classifiers presume the influence on a certain class of a variable values. The Naïve Bayes inducer calculates the classes conditional probabilities provided the case, and a class with highest posterior is selected. Naïve Bayes is a classifier focused on probability, which measures the probabilities of the specified instances. The likelihood set shall be determined by computing the data sets frequency and combinational vales. The class likelihood is chosen by the classifier and is near to the rear end. The Naïve Bayes is a classifier with multiclass that operates well with a controlled approach to learning.

2.2 EXTRA TREE CLASSIFIER

Extra Tree Classifier is also known as Extremely Randomized Trees Classifier. This classifier is similar to random Forest Classifier. This classifier is highly Randomized Trees Classifier is a form of learning technique for the ensemble that aggregates the effects of several decision trees obtained in "forest" to generate the outcome

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

of classification. It is quite close in principle to this classifier, and varies only from it in a way decision trees are designed in the forest. Every stump is built with all the available data in the training data. For the formation of root node or any of the other nodes, it is scheduled by finding best split by the method of searching the randomly selected subset. The splitting of every chosen element is picked at random. The average decision-stump size is one.

2.3 GRADIENT BOOSTING CLASSIFIER

Boosting is a process by which poor learners are transformed into good learners. Growing new tree is a match to a updated version of original data set while boosting. It is best to clarify the gbm(Gradient Boosting Method) by first implementing the Ada. The Adaboost starts with decision tree training in which each measurement is given an equivalent weight. We increase the weights of the observations after the evaluation of the first tree which has a difficulty factor for classifying and by lowering the weights which seems easier to classify. Therefore on this weighted info the second tree is established. Gradient boosting is a method used for both classification and regression in machine learning. This generates a predictive model in the context of a low predictive model ensemble. Boosting is identical, but sample size is rendered smarter. Subsequently we are adding more and more weight to the results hard to define.

2.4 COUNT VECTORIZER

The Count Vectorizer offers an easy way to both tokenize set of text files and create a corpus of established terms but also to encrypt new files using that vocabulary.

2.5 ADA BOOSTER

Ada-Boost is a classifier similar to Random Forest Classifier and is also an ensemble classifier. A classifier is a classifier with multiple classifier algorithms where the output is the combination result of the output of all those remaining classifiers. This classifier identifies the weak classifiers, by identifying it combines the weak classifier in order to form a strong classifier from a weak classifier. But if we mix several classifiers with training collection set at each iteration and give the correct weight in the final vote, we can have strong overall classifier accuracy ranking. This classifier choose the set of training based on the efficiency of its previous training. The trained classifier weight age is dependent on the efficiency achieved. By using the random subset of all the training sets the weak classifier is trained. But the random set id not 100% considered as random. Weight is assigned to each and every training item at every level.

The item that is misclassified is assigned with highest weight as appears with highest probability in the next classifier of the training subset. Since after the every classifier is trained accordingly, weights are assigned to the classifiers always based on their accuracy. The classifier with highest weight in order it can get high impact in output. Decision tress are the weak classifiers in this classifier called as decision stumps. Boosting is a ensemble approach that uses a variety of weak classifiers to construct a powerful classifier. This is achieved by using training data to construct a model and then developing a second model which attempts to correct the first model errors. Models are introduced when the training set is correctly estimated or until a sufficient number of models is included. AdaBoost classifier was the first very effective binary-classification boosting algorithm created. It is the best starting point for boosting your understanding.

2.6 IMAP

IMAP is known as Internet Message Access Protocol. Whenever a mail is opened it is connected to servers and the data is released. So it is possible to check our mail from any other servers. Email servers can always be used for either sending and receiving email messages. In our project a email id with its password is provided. A list of upto 10 mails are collected and classification is done on those emails.

III. RESULTS

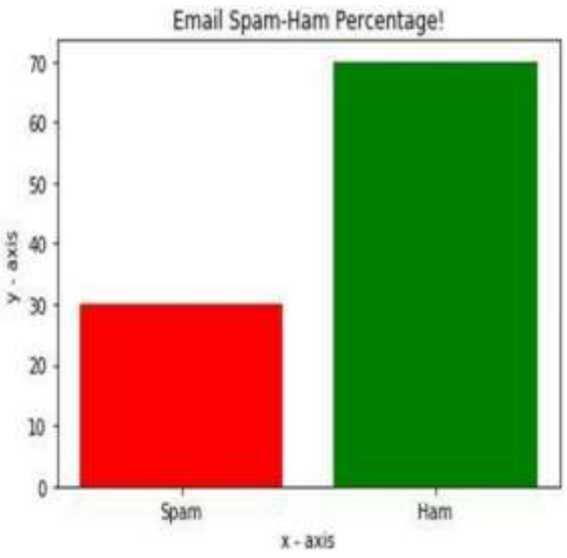

Figure 2 : Graph representation

By using Imap we need to give access to less secure apps without giving Access it cannot get mails from the required mails, so security is one of the major drawback and also hybrid algorithms takes lot of time for training the model and for classification purpose

.

IV. CONCLUSION

In this paper, many machine learning techniques for classification and regression problems. The strategy of these algorithms is to identify spam and ham emails. Spam email is one of the most promising internet issues in the world. Each and every technique is applied and accuracy is observed. The algorithms with highest accuracy are considered and later applied for hybrid algorithms.

The future direction of spam email detection is a simulation of all the factors that can affect email accounts with more number of emails receiving from unknown person. By this mostly we can eradicate the spam emails

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

without entering into our inbox so that we don’t have to clean our inbox regularly and we can know which email belongs to ham or spam and also we can classify which spam email belongs to which sector.

REFERENCES

1. Bhowmick, Alexy & Hazarika, Shyamanta. (2018). E-Mail Spam Filtering: A Review of Techniques and Trends. 10.1007/978-981-10- 4765-7_61.
2. Machine learning for email spam filtering: review, approaches and open research problems Emmanuel Gbenga Dada a,* , Joseph Stephen Bassi a , Haruna Chiroma b , Shafi'i Muhammad Abdulhamid c , Adebayo Olusola Adetunmbi d , Opeyemi , K., Otsuka, S., Apip, and Saito, K. (2016).
3. G.H. Schmitz, J. Cullmann, “PAI–OFF: A new proposal for email spam filtering ,”
4. Ann Nosseir , Khaled Nagati and Islam Taj- Eddin, “Intelligent Word-Based Spam Filter Detection Using MultiNeural Networks”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.
5. "Email Statistics Report, 2014-2018." Email Statistics Report, 2014-2018. Ed. Sara Radicati. The Radicati Group, Inc., 14 Apr 2014. Web. 25 Apr. 2014.
6. Padmaja P and B. Lakshmi Ramani. “ Adaptive Fuzzy System with Robust GSCA-based Fuzzy Rule Extraction for Data Classification,” Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 01-Special Issue, 2018.
7. Tumuluru, P. and Ravi, B. GOA-based DBN: Grasshopper Optimization Algorithm-based Deep Belief Neural Networks for Cancer Classification. International Journal of Applied Engineering Research 12 (24) (2017).
8. Tumuluru, P. and Ravi, B. Chronological Grasshopper Optimization Algorithm- based Gene Selection and Cancer Classification. Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, No. 3, 2018.
9. Tumuluru P, Bhramaramba R, "A Framework for Identifying of Gene to Gene Mutation causing Lung Cancer using SPI - Network", International Journal of Computer Applications, vol. 152, no. 10, pp. 21-26, October 2016.
10. Praveen Tumuluru, et al. "Credentials of Lung-Cancer Associated Genes Using Protein-Protein Interaction Network", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, No. 3, pp. 82-89, March 2016.
11. Tumuluru P, Bhramaramba Ravi "Dijkstra’s based Identification of Lung Cancer Related Genes using PPI Networks", International Journal of Computer Applications (0975 – 8887), Vol. 163, No. 10, pp. 1-10, April 2017.
12. Tumuluru P, Bhramaramba R "A Survey on Gene Expression Classification Systems", International Journal of Scientific Research and Review ISSN NO: 2279-543X, Volume 6, Issue 12, 2017.
13. Tumuluru P, Burra Lakshmi Ramani et al. " OpenCV Algorithms for facial recognition", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.
14. Burra Lakshmi Ramani, Tumuluru P et al. “ Deep Learning and Fuzzy Rule-Based Hybrid Fusion Model for Data Classification” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-

3878, Volume-8 Issue-2, July 2019.

15. Tumuluru P, Radha Manohar Jonnalagadda et al. “Extreme Learning Model Based Phishing Classifier” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.
16. Dr. Padmaja P and B. Lakshmi Ramani, “Adaptive Lion Fuzzy System to Generate the Classification Rules using Membership Functions based on Uniform Distribution” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14421-14433.
17. Tumuluru, P., Lakshmi, C.P., Sahaja, T., Prazna, R. “A Review of Machine Learning Techniques for Breast Cancer Diagnosis in Medical Applications “Proceedings of the 3rd International Conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019.
18. Nalajala, S., Akhil, K., Sai, V., Shekhar, D.C., Tumuluru, P. “Light Weight Secure Data Sharing Scheme for Mobile Cloud Computing” Proceedings of the 3rd International Conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019.

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020