JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ENHANCED TOOL FOR SPAM ANALYSIS: SPAM CLASSIFIER

[1]Anuj Mourya, [2]Adarsh Arya, [3] Dr. Nidhi Saxena

[1]Undergraduate, [2] Undergraduate, [3]Assistant Professor

[1,2,3]CSE Department,

[1,2,3]SRMCEM, Lucknow, India

[1]anujmaurya704@gmail.com, [2]adarsh1arya@gmail.com, [3]nidhi.shivansh@gmail.com

Abstract: A better mean of digital communications in the internet world is the spam email, which can sent to a group of individuals or a company. These spam emails causes a serious threat to the user. Another method of spamming is by creating a temporary email register which receive emails that can be deleted after some certain amount of time. This method is contemporary used for misusing those temporary email addresses for sending free spam emails without revealing spammers real account details. These attacks create major problems like lack of storage, etc. Hence it is important to introduce an efficient detection mechanism through extraction and classification which detect spam emails. This can be used through a novel NLP Algorithm which is based on Multinomial Naive Bayes approach. With its help the proposed approach will reduce the spam emails, method improves the accuracy of spam email filtering, as the use of NLP makes the system to detect Multinomial Naive Bayes approach uses multiple decision trees and uses a random node for filtering the spams.

Keywords—Vectorization, Stop words, Tokenization, Adding Corpus

I. INTRODUCTION

Recently the commercial or the bulk e-mail which is known as spam, is now a big trouble over the internet. Spam is waste of time, storage memory and bandwidth of the communication. The problem of spam e-mail is increasing for years. In recent statistics, half of all emails are spam which about more than 15 billion email everyday which cost internet users a huge amount of loss. Automatic e-mail filtering is the most effective method for countering spam at the moment which is a hard competition between spammers and spam-filtering methods which is going on. Several years ago most of the spam are dealt by blocking e-mails coming from unknown addresses or filtering of the messages with certain subject lines. Spammers are using several tricky methods to overcome the filtering methods like using random sender addresses and/or append random characters to the beginning or the end of the message subject line [11]. Knowledge engineering and ML are the two general approaches used in spam classification. In knowledge engineering approach a set of rules has to be specified in which emails are categorized as spam or ham. A set of such rules should be created by the user of by the filter. By applying this method, no promising results shows because the rules are constantly updating and maintained, which is a waste of time and it is not convenient for most users. ML approach is more efficient than knowledge engineering approach; it does not require specifying any rules [4]. As, a set of training samples, these samples belong to a set of pre classified e-mail messages. A specific algorithm is then used to learn the classification rules from these e-mail messages. Machine learning approach has been widely studied and there are lots of algorithms can be used in spam classification. This include several algorithms such as Naïve Bayes, support vector machines, Neural Networks, K-nearest neighbor, rough sets and the artificial immune system.

II. LITERATURE SURVEY

Spam mail, known as un-solicited bulk e-mail or junk mail which is sent as a group of recipients who have not requested for it. The aim of spam classifier is governed by un-solicited e-mails which automatically request from a user's mail model. These models have already caused many problems such as filling mailboxes, mail, consuming users time and energy to sort through it, Spam mail, known as un-solicited bulk e-mail or junk mail which is sent as a group of recipients who have not requested for it. The aim of spam classifier is governed by un-solicited e-mails which automatically request from a user's mail model. These models have already caused many problems such as filling mailboxes, mail, consuming users time and energy to sort through it, According to a series of surveys conducted by CAUBE.AU 1, the number of total spams received by 40 % email addresses has increased by a factor of six in two years (from 1753 spams in 2000 to 10,847 spams in 2001) [4]. Therefore, it is very hard to develop spam filters that effectively remove the growing volumes of unwanted mails automatically before they enter a user's mailbox.

D. Puniskis [5] in his research applied the neural network approach to the classification of spam. His method are different from others as he employs attributes composed of descriptive characteristics of the evasive patterns that spammers employ rather than

using the context or frequency of keywords in the message. The data used is corpus of over 2700 legitimate and over 1800 spam emails received during period of several months of the same year. The result shows that ANN is good but in spite of being good it's not suitable for using alone as a spam filtering tool.

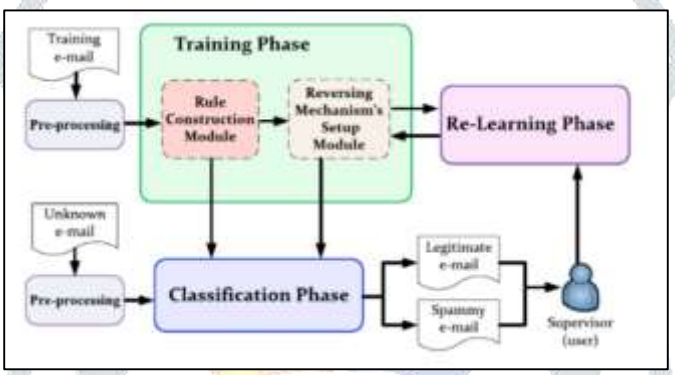
In [6] email data was classified using four different classifiers (Neural Network, SVM classifier, Naïve Bayesian Classifier, and J48 classifier). The experiment was performed based on different data size and different feature size. The final classification result should be '1' if it is finally spam, otherwise, which make a binary tree, could be efficient for the dataset which could be classified as binary tree.

III. OBJECTIVES

The main objective of this study is to

- To detect unsolicited, unwanted, and virus-infested email(called spam) and stop it from getting into email inboxes.
- To prevent users from fraud mail.
- This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
- It is achieved by creating user-friendly screens for the data entry to handle large volume of data.
- Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus, the objective of input design is to create can input layout that is easy to follow.

IV. PROPOSED SYSTEM



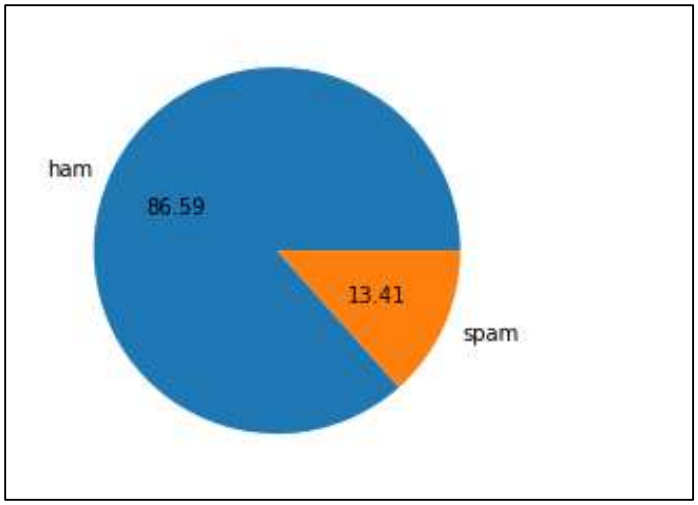
For analyzing real time dataset and to predict the performance, the supervised learning algorithms were adopted here. Different algorithms use different biases for generalizing different representations of the knowledge. Therefore, they tend to error on different parts of the instance space. The combined use of different algorithms could lead to the correction of the individual uncorrelated errors. There are two main paradigms for handling an ensemble of different classification algorithms: Classifier Selection and Classifier Fusion. The first one selects a single algorithm for classifying new instances, while the latter fuses the decisions of all algorithms. This section presents the most important methods from both categories. Classifier Selection is a very simple method, which produces Selection or Select Best. This method evaluates each Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text .

4.1. Classification algorithm used

Naive Bayes-classifier:
Naive Bayes-classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". The Naive-Bayes inducer computes conditional probabilities of the classes given the instance and picks the class with the highest posterior. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. The basic concept of it is to find whether an e-mail is spam or not by looking at which words are found in the message and which words are absent from it. Naïve Bayes classifies can handle an arbitrary number of independent variables whether continuous or categorical. Given a set of variables, X={X1, X2…..Xd}, we can construct the posterior probability for the event Cj among a set of possible outcomes C = {c1, c2……cd}

$$P(y|x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

Using Bayes rule, we can label the new case with a class Cj that achieves the highest posterior probability.

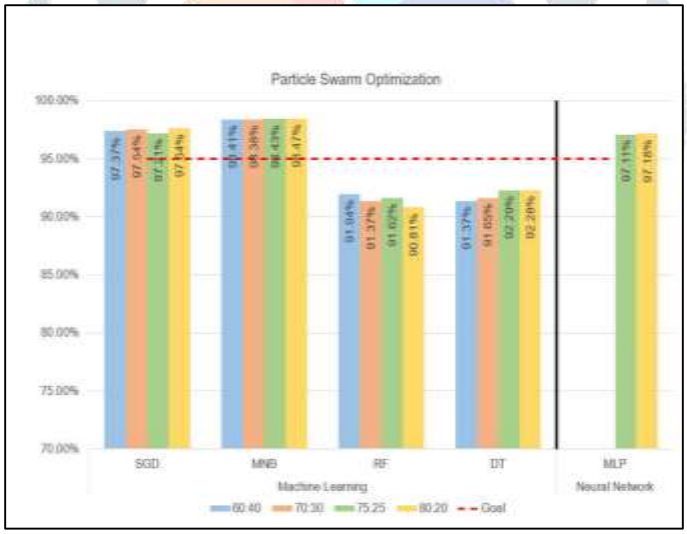


This dataset is collected from Kaggel. This pie-chart shows that that dataset is imbalanced due to which machine train well. So, the result shows the better accuracy.

4.2 Advantages of Naïve Bayes

- It doesn't require as much training data
- It handles both continuous and discrete data
- It is highly scalable with the number of predictors and data points
- It is fast and can be used to make real-time predictions
- It is not sensitive to irrelevant features

V. RESULT EVALUATION



In the above figure we have showed different algorithms such as

5.1 Stochastic Gradient Descent (SGD) in which in Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. In the process of Gradient Descent, there is a term known as 'batch' which describes the total number of samples which from a dataset used for calculating the gradient for each iteration. In typical Gradient Descent optimization, like Batch Gradient Descent, the batch is taken to be the whole dataset. As, the whole dataset is useful for getting to a less noisy and less random manner, but the problem arises when our datasets get big. This algorithm is showing an accuracy of about 97%.

5.2 Multinomial Naive Bayes (MNB) The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program predict the tag of a text, such as an email, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance. The Naive Bayes classifier is of a number of algorithms that all have one thing in common in which each feature being classed is unrelated to any other feature. A feature's existence or absence has no bearing on the inclusion or exclusion of another feature. This algorithm is showing an accuracy of about 98%. Thus we are using this algorithm for our Model.

5.3 Random Forest (RF) Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The Accuracy of this algorithm is showing 90%.

Decision Tree (DT) Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The Accuracy of this algorithm is showing 92% for our model.

5.4 Multilayer perceptron (MLP) A multilayer perceptron strives to remember patterns in sequential data, because of this, it requires a "large" number of parameters to process multidimensional data. It is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP. MLPs with one hidden layer are capable of approximating any continuous function. The Accuracy of MLP is showing 97% for our model.

VI. MEASURING THE PERFORMANCE

The meaning of a good classifier varies depending on the domain for which it is used. For example, in spam classification it is very important not to classify legitimate messages as spam as it can lead to. e.g. economic or emotional suffering for the user.

In above all algorithm we found that Naïve Bayes algorithm is better than all of the others. The performance of the other algorithms that has been applied on the dataset with 80:20 split s given below:

Classifier	Split Set			
	60:40	70:30	75:25	80:20
SGD	97.37%	97.54%	97.21%	97.64%
MNB	98.41%	98.38%	98.43%	98.47%
RF	91.94%	91.37%	91.62%	90.81%
DT	91.37%	91.65%	92.29%	92.28%
MLP	-	-	97.11%	97.18%

Spam emails are one of the critical issues over internet in recent communication systems. Spammers send spam emails by misusing the facilities in the online communication by various methods, including sending spam emails by using temporary email addresses that affects the users and involved organizations. In this paper, a method of Natural Language Processing based on Naive Bayes-classifier is proposed which helps in easily detecting spam emails and temporary email addresses effectively that violates the privacy of users and prevents exposing private data of the users. This method can enhance the privacy of email sender and recipients and reduces security risks and in future the work is subjected to get better progress and accuracy by boosting the dataset for better features and classification systems.

VII. CONCUSION AND FUTURE SCOPE

Spam emails are one of the critical issues over internet in recent communication systems. Spammers send spam emails by misusing the facilities in the online communication by various methods, including sending spam emails by using temporary email addresses that affects the users and involved organizations. In this paper, a method of Natural Language Processing based on Naive Bayes-classifier is proposed which helps in easily detecting spam emails and temporary email addresses effectively that violates the privacy of users and prevents exposing private data of the users. This method can enhance the privacy of email sender and recipients and reduces security risks and in future the work is subjected to get better progress and accuracy by boosting the dataset for better features and classification systems.

VIII. ACKNOWLEDGMENT

We would like to thank all the working staff of Shri Ramswaroop Memorial Group of Professional College of Engineering and Management and a great thank to our project guide Dr. Nidhi Saxena who gave her best to make us do this great project. Moreover, thanks to all the authors whose papers and books we referred during this project. Without help of all these resources, we would have never completed this project.

REFERENCES

[1] Androutsopoulos, J. Koutsias, "An evaluation of naïveBayesian anti-spam filtering", 11thEuropean Conference on Machine Learning (ECML 2000),pp9–17, 2000.

[2] K. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering", 10th Conference of the European Chapter of the Association for Computational Linguistics 2003.

[3] C. Miller, "Neural Network-based Antispam Heuristics", Symantec Enterprise Security (2011), www.symantec.com Retrieved December 28, 2011

[4] Esha Bansal and Anupam Bhatia, "Support Vector Machine for Multiclass Handwritten Digits" International conference on Advanced Information Communication Technology in Engineering (ICAICTE-2K13).

[5] AnirudhHarisinghaney, Aman Dixit, Saurabh Gupta, and AnujaArora , "Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN Algorithm, " ICROIT 2014, India, Feb 6-8 2014.
[6] Hammad, A.S.; El-Halees, A. An Approach for Detecting Spam in Arabic Opinion Reviews. Ph.D. Dissertation, Islamic University of Gaza, Gaza Strip, Palestine, 2013.
[7] Malika Ben Khalifa, Zied Elouedi, Eric Lefevre,"An Evidential Spammer Detection based on the Suspicious Behaviors' Indicators.", Auckland University of Technology. August 11,2020.
[8] Nikhil Kumar, Sanket Sonowal, Nishant, "Email Spam Detection Using Machine Learning Algorithms.", Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.
[9] Priyanka Verma, Anjali Goyal and Yogita Gigras, "Emai phishing: Text classification using natural language processing.", Computer Science and Information Technologies, Vol. 1, No. 1, May 2020, pp. 1~12, ISSN: 2722-3221, DOI: 10.11591/csit.v1i1.p1-12.
[10] Li, J.; Ott, M.; Cardie, C.; Hovy, E. Towards a general rule for identifying deceptive opinion spam. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014.