E-mail Spam Classification Using Naïve Bayesian Classifier

Priyanka Sao, Pro. Kare Prashanthi

***Abstract*— E-mail spam is the very recent problem for every individual. The e-mail spam is nothing it's an advertisement of any company/product or any kind of virus which is receiving by the email client mailbox without any notification. To solve this problem the different spam filtering technique is used. The spam filtering techniques are used to protect our mailbox for spam mails. In this project, we are using the Naïve Bayesian Classifier for spam classification. The Naïve Bayesian Classifier is very simple and efficient method for spam classification. Here we are using the Lingspam dataset for classification of spam and non-spam mails. The feature extraction technique is used to extract the feature. The result is to increase the accuracy of the system.**

***Index Terms*— E-mail spam, Classification, Feature Extraction, Naïve Bayesian Classifier.**

I. INTRODUCTION

Email spam is operations which are sending the undesirable messages to different email client. The historical backdrop of Email spam is begin before 2004, however these are the enormous parts that convey spam to the way it is today. Commercialization of the web and united as complete thing of electronic post as a ready to be got to method for news has another face things coming in of not needed data and sends messages

One subset of UBE is UCE (Unsolicited Commercial Email). The inverse of "spam", email which one needs, is called "ham", normally when alluding to a messages mechanized examination, (for example, Bayesian Filtering).

Email spam targets singular clients with regular postal mail messages. Email spam records are regularly made by checking Usenet postings, taking Internet mailing records, or scanning the Web for locations. Email spams normally cost clients cash out-of-pocket to get. Numerous individuals - anybody with measured telephone administration - read or get their mail while the meter is running, as it were. Spam costs them extra cash. On top of that, it costs cash for ISPs and online administrations to transmit spam, and these expenses are transmitted specifically to endorsers.

Progressively, email spam today is sent by means of "zombie systems", systems of infection or worm-contaminated PCs in homes and workplaces around the world. Numerous advanced worms introduce an indirect access that permits the spammer to get to the PC and utilization it for pernicious purposes. This entangles endeavors to control the spread of spam, as a rule the spam does not clearly start from the spammer. In November 2008 an ISP, McColo, which was giving support of botnet administrators, was depeered and spam dropped 50 to 75 percent all inclusive. In the meantime, it is turning out to be clear that malware creators, spammers, and phishers are gaining from one another, and perhaps framing different sorts of organizations.

There are two fundamental sorts of spam, and they have distinctive consequences for Internet clients. Cancellable Usenet spam is a solitary message sent to 20 or more Usenet newsgroups. (Through long experience, Usenet clients have found that any message presented on such a large number of newsgroups is frequently not significant to most or every one of them.) Usenet spam is gone for "prowlers", individuals who read newsgroups yet seldom or never post and dole their location out. Usenet spam denies clients of the utility of the newsgroups by overpowering them with a torrent of promoting or other unessential posts. Moreover, Usenet spam subverts the capacity of framework directors and proprietors to deal with the themes they acknowledge on their frameworks.

Email Filtering/Spam Filtering:

To detect unsolicited and unwanted email and prevent those unwanted messages from getting to a users inbox is called spam filter. The spam filter is a program like other types of filtering program looks for certain criteria on which it bases judgments.

The input of email filtering software is emails. The message through unchanged for delivery to the user's mailbox is the output of email filter. Some of the mail filters are able to edit messages during processing.

Mail filters have differing degrees of configurability. Once in a while they settle on choices taking into account coordinating a consistent expression. Different times, essential words in the message body are utilized, or maybe the email location of the sender of the message. Some more propelled channels, especially hostile to spam channels, use measurable archive order methods, for example, the guileless Bayes classifier. Picture sifting can likewise be utilized that utilization complex picture examination calculations to identify skin-tones and particular body shapes typically connected with obscene pictures.

Mail filters can be introduced by the client, either as independent projects (see interfaces underneath), or as a major aspect of their email project (email customer).

In email programs, clients can make individual, "manual" channels that then naturally channel mail as indicated by the picked criteria. Most email projects now likewise have a programmed spam separating capacity. Network access suppliers can likewise introduce mail channels in their mail exchange operators as a support of the greater part of their clients. Because of the developing danger of fake sites, Internet administration suppliers channel URLs in email messages to uproot the risk before clients click.

Normal uses for mail filters incorporate arranging incoming email and evacuation of spam and PC infections. A less basic utilization is to investigate active email at a few organizations to guarantee that workers consent to proper laws. Clients may additionally utilize a mail filter to organize messages, and to sort them into organizers in light of topic or other criteria.

II. RELATED WORK

Web spam which is a major issue throughout today's web search tool; consequently it is important for web crawlers to have the capacity to detect web spam amid creeping. The Classification Models are designed by machine learning order algorithm. [2] The one machine learning algorithm is Naïve Bayesian Classifier which is also used in [1] to separate the spam and non-spam mails. Big Data analyzing framework which is also outline for spam detection. Extricating the feeling from a message is a method for get the valuable data. In Machine learning innovations can gain from the preparation datasets furthermore anticipate the choice making framework hence they are broadly utilized as a part of feeling order with the exceptionally precision of framework. [3]

Most of the research work has already been carried out on improving the efficiency and accuracy of Naïve Bayesian approach. Paul Graham's Naïve Bayesian Machine learning approach is used to improve the efficiency of Bayesian approach. [1] for vast dataset also using the naïve Bayesian algorithm and increment the precision of NBC. [4] The research work has also carried out for increase the accuracy and time efficiency of system.

III. PROBLEM IDENTIFICATION

Email Spam is most crucial matter in a social network. There are many problem created through spam. The spam is nothing this is unwanted message or mail which the end user doesn't want in our mail box. Because of these spam the performance of the system can be degraded and also affected the accuracy of the system. To send the unsolicited or unwanted messages which are also called spam is used in Electronic spamming. In this project explain about the email spam, where how spam can spoil the performance of mailing system. In the previous study there are many types of spam classifier are present too detect the spam and non-spam mails.

There are different email filtering techniques are also used in spam detection. Mostly popular filters or classifier are: Decision tree classifier, Negative Selection Algorithm, Genetic Algorithm Support Vector Machine Classifier, Bayesian Classifier etc. From the previous study we identify that Support Vector Machine (SVM Classifier) are used for email spam classification. But it takes very much time for detecting spam. The SVM Classifier has also wrongly classified the messages. So the system can be on a risk. The error rate of SVM Classifier is very high. In this project there is also discussion in the Feature Selection process. There are different feature extractions techniques are present which are used in extracting the messages.

Solution of the Problem:

To solve the problem of previous study in this project we are uses the Naive Bayesian Classifier for classify the spam and non-spam mails. The naive Bayesian Classifier is one of the most popular and simplest methods for classification. Naive Bayesian Classifiers are highly scalable, learning problem the number of features are required for the number of linear parameter. Training of the large data simple can be easily done with Naive Bayesian Classifier, which takes a very less time as compared to other classifier. The accuracy of system is increase using Naïve Bayesian Classifier.

IV. METHODOLOGY

E-mail spam classification has major issue in today's electronic world. To solve this problem the different spam classification methods are used. Using this spam detection technique we can identifies the spam and non-spam mails in our mailbox. In this work we are using the Naïve Bayesian Classifier for email spam classification.

In this work also use feature extraction techniques for providing efficient dataset. The feature extraction techniques are used when the input data is too large and it is redundant in nature so feature is extracted to obtain an accurate result. In this work we are using the word-count algorithm for extracting feature from the dataset. Here we use the Lingspam data set which contains total 960 mails in which 700 are train dataset and 260 are test dataset. The train and test data are further divided in two parts spam mails and non-spam i.e. 50% of train dataset are spam dataset and 50% are non-spam dataset as same for the test dataset.

The Feature Extraction:

The word-count algorithm is very simple to implement and provide a flexible result. In this algorithm we pre-process the dataset and remove the stop-words and non-words in dataset. And then it counts the total number of unique word out of the total word and finds the frequency of that word in a particular document. The main thing about this algorithm is to makes a dictionary. In that dictionary the path of the file is stored which is pre-processed. So the redundancy problem is removed. For counting the word and store the frequency of that word is very helpful to find the unique word.

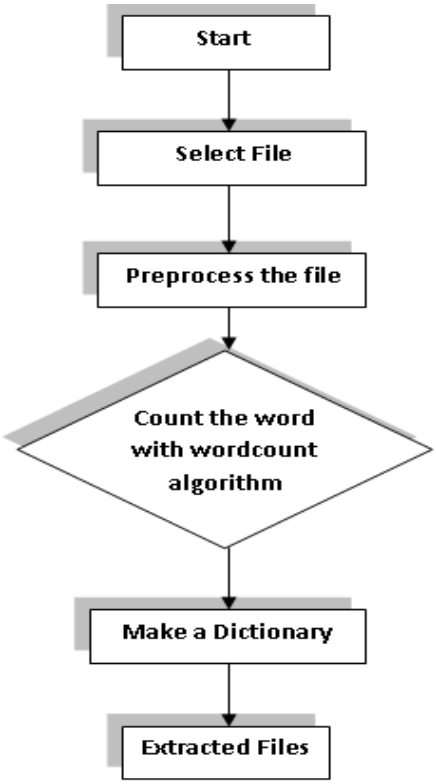

Figure 3.1 Feature Extraction Method

Algorithm:

Step 1: Select the file from the dataset.
Step 2: Pre-process the file and removing the stop-word.
Step3: Count the total word of the file and find the unique word of that file.
Step 4: Calculate the frequency of words.
Step 5: Make a dictionary and store the file path.
Step 6: Extracted Feature.

Description: In this feature extraction word-count algorithm in which three steps are present they are pre-processing, count the word, make dictionary. The first is to select the file from the dataset. Then second pre-process the data in which first remove the stop word and non-words from the document. The third step for feature extraction is to count the unique word from total number of words. So we can calculate the frequency of that word in a document. The forth step is to make a dictionary and store the path of document this can solve the redundancy problem. The extracted data are received after all steps are complete.

The proposed methodology:

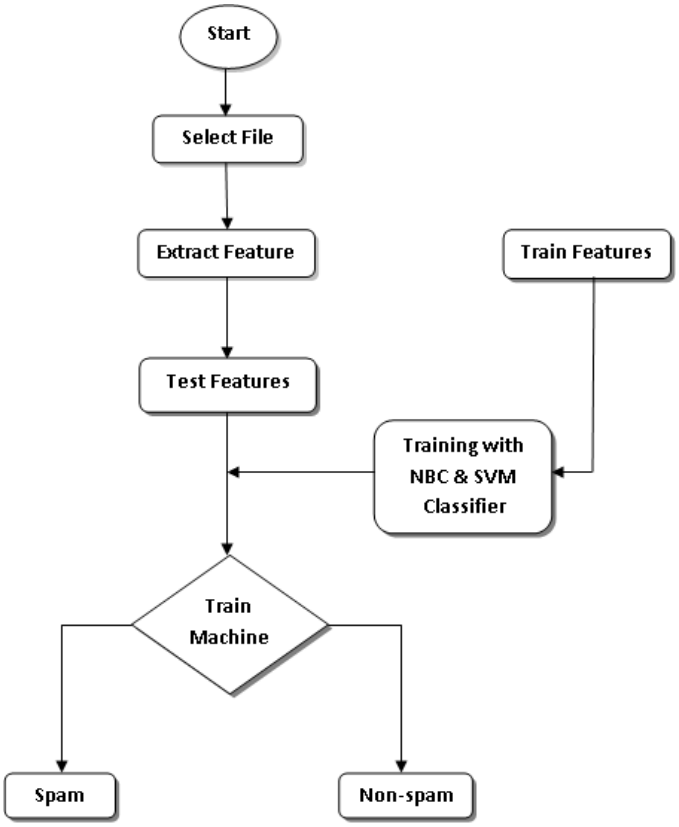

Figure: The Proposed Methodology
Algorithm:

Step 1: Select the file
Step2: Extracting the feature with help of wordcount algorithm.
Step 3: Training the dataset with the help of Naive Bayesian Classifier.
Step 4: Find the probability of spam and non-spam mails.
Prob_spam = (sum(train_matrix(spam_indices,)) + 1) ./ (spam_wc + numtokens)
Prob_nonspam = (sum(train_matrix(nonspam_indices,)) + 1) ./ (nonspam_wc + numtokens)
Step 5: Testing the dataset
log_a = test_matrix*(log(prob_tokens_spam))' + log(prob_spam)
log_b = test_matrix*(log(prob_tokens_nonspam))'+ log(1 - prob_spam)
if
output = log_a > log_b
then document are spam
else the document are non-spam
Step 6: Classify the spam and non-spam mails.
Step 7: compute the error of the text data and calculate the word which is wrongly classified
Numdocs_wrong = sum(xor(output, text_lables))
Step 8: display the error rate of text data and calculate the fraction of wrongly classified word
Fraction_wrong = numdocs_wrong/numtest_docs

Description: In this work we are describing the method which is used to perform e-mail spam classification. The first step is to select the file from the dataset and apply the feature extraction technique for extracted feature. For which we are using the word-count algorithm. The next step is training the dataset which are extracted by the feature extraction technique. For training the data we can calculate the probability of spam and non-spam words in the document. The next step is to test the data with the help of Naïve Bayesian Classifier for which calculation the probability of spam and non-spam mails and make a prediction which value is higher. If spam words are greater than non-spam words in a mail then the mail is spam mails otherwise non-spam mails.

In the next step we are calculating the words which are wrongly classified by the classifier and calculate accuracy of the classifier and also calculate the error rate of classifier by calculating the fraction of word which is wrongly classified and total number of words in document.

V. RESULT & DISCUSSION

In this project work we are explain about the e-mail spam classification to identify the spam and non-spam mails. For this purpose we are using Naïve Bayesian Classifier. In this project we are creating an email spam classification system for classify the spam and non-spam mails. For this we are taking the Lingspam dataset to run this experiment. In a Lingspam dataset we are taking total 960 mails in which 700 train dataset and 260 test dataset. Out of 700 train dataset the 350 are spam mails and 350 are non-spam mails. Similarly the 260 test dataset is containing 130 spam mails and 130 non-spam mails.

Here we are present different reading for all four trained dataset which are tested by the classifier i.e. Naive Bayesian Classifier and Support Vector Machine. Hence shown the different readings and calculation of result:

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

Table 4.1 Reading of different classifier

This reading contains the text data which are classify by the classifier and provide the word which are wrongly classified and error rate of classifier. Hence we can show the overall result which are provided by classifier. And say that the Naive Bayesian Classifier classifies mostly word in accurate way. When the number of dataset is increase the Naive Bayesian Classifier produce a better result as compared to Support Vector Machine.

Hence we can show the graph for display the accuracy and error rate of both classifier (Naive Bayesian Classifier and Support Vector Machine).

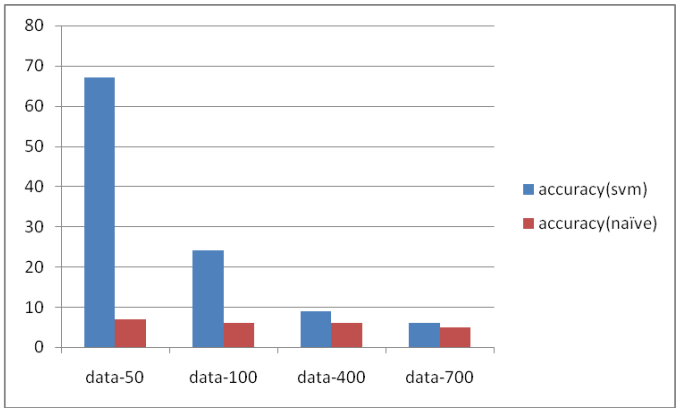

Figure 4.1: representing the accuracy graph

In this graph we can show the accuracy of classifiers (Naive Bayesian Classifier and Support Vector Machine). For this we are calculating the words which are wrongly classified by the classifier. The classifier which are classified more numbers of wrong words is less accurate as compared to the classifier which classified less numbers of wrong word. So here the Naive Bayesian Classifier is more accurate the Support Vector Machine.

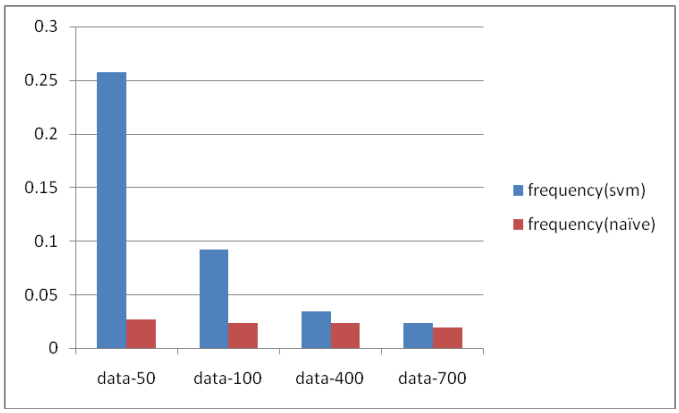

Figure 4.2: representing the error rate of classifier

In this graph we show the error rate of classifiers for which we calculate the fraction of a word which are wrongly classified into the total number of words. Hence we can see that the error rate of blue bar is greater than the error rate of red bar, so we can say the Support Vector Machine provide the high error rate than the Naive Bayesian Classifier.

VI. CONCLUSION

Spam is a big problem of today's world; to solve this problem the spam classification system is created to identify the spam and non-spam mails. The spam messages are the unwanted messages which the end user clients are receiving in our daily life. Spam mails are nothing it is the advertisement of any company, any kind of virus etc.

To solve this problem create an email spam classification system and identifies the spam and non-spam mails.

Here we are using the Naïve Bayesian Classifier and extracting the word using word-count algorithm. After calculation we find that naïve Bayesian classifier has more accurate the support vector machine. The error rate is very low when we are using the Naïve Bayesian Classifier. So we can say that Naïve Bayesian Classifier produce better result than Support Vector Machine.

REFERENCES

[1] Sharma K. and Jatana N. (2014)"Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach" IEEE 2014 pp. 939-942.

[2] Sharma A. and Anchal (2014), "SMS Spam Detection Using Neural Network Classifier",ISSN: 2277 128X Volume 4, Issue 6, June 2014, pp. 240-244.

[3] Ali M. et al (2014), , "Multiple Classifications for Detecting Spam email by Novel Consultation Algorithm", CCECE 2014, IEEE 2014, pp. 1-5.

[4] Liu B. et al (2013) "Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier" IEEE 2013 pp.99-104.

[5] Belkebir R. and Guessoum A. (2013), "A Hybrid BSO-Chi2-SVM Approach to Arabic Text Categorization", IEEE 2013, pp. 978-984.

[6] Blasch E. et al (2013), Kohler, "Information fusion in a cloud-enabled environment," High Performance Semantic Cloud Auditing, Springer Publishing.

[7] Allias N. (2013) "A Hybrid Gini PSO-SVM Feature Selection: An Empirical Study of Population Sizes on Different Classifier" pp 107-110.

[8] Prasad N. et al (2013) "Comparison of Back Propagation and Resilient Propagation Algorithm for Spam Classification", Fifth International Conference on Computational Intelligence, Modelling and Simulation, IEEE 2013, pp. 29-34.

[9] Jia Z. et al (2012) "Research on Web Spam Detection Base on Support Vector Machine" IEEE 2012 pp. 517-520.

[10] Panigrahi P. (2012) , "A Comparative Study of Supervised Machine Learning Techniques for Spam E-Mail Filtering", Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012, pp. 506-512.

[11] Clark J. et al (2010), "A Neural Network Based Approach to Automated Email Classification.

[12] Sun X. et al (2009), "Using LPP and LS-SVM For Spam Filtering", School of Information Science and Engineering Henan University of Technology IEEE 2009, pp. 4244-4246.

[13] Hmeidi I. and Hawashin B. (2008),"Performance of KNN and SVM classifiers on full word Arabic articles," Advanced Engineering Informatics, vol. 22, no. 1, pp. 106-111

Website:

1. http://www.Anti-spamtechniquesWikipedia.org
2. http://www.EmailspamWikipedia.org
3. http://www.TextcategorizationScholarpedia.com

Priyanka Sao

Mtech Scholar from Computer Science & Engineering specialised in Software Engineering from Rungta College of Engineerin & Technology Bhilai (C.G.), done BE in Information Technology from Central College of Engineerin & Management Raipur (C. G.)

Miss Kare Prashanthi

Working as Assistant Professor in Computer Science & Engineering department, RCET Bhilai has 2.5 years of teaching experience with 1 year in industory. She has published 2 papers in international, 2 papers in national conferences and 1 paper in international journal. She is M-Tech(hons) from VTU, Karnataka in Computer Network Engineering.

