International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.01, April 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

A Survey on understanding of Spam & algorithms of Data mining as Spam filtering Techniques

Sachidanand Chaturvedi[1] , RavindraGupta[2] ,Dr,Varsha Namdev[3]

[1]Second Year Student Of M.Tech. (Software Engineering),Department of Computer Science & Engineering ,SarvepalliRadhakrishnan University Bhopal (India)

[2]Asst.Prof. , Department of Computer Science & Engineering , SarvepalliRadhakrishnan University Bhopal (India)

[3]HOD , Department of Computer Science & Engineering Department , SarvepalliRadhakrishnan University Bhopal (India)

ABSTRACT

Spam mail or junk mail is one of the evolving problem in today's internet world which directly and indirectly causes financial harm. this paper provides a survey on data mining techniques performed for the spam filtering. This includes both supervised and unsupervised model of machine learning. . Existence of spam mail causes the unnecessary network traffic, wastage of memory, time and bandwidth. Finally, it causes direct or indirect financial loss or damage. It sometimes contains harmful and malicious virus that get activated after one click at its link and corrupt our system. Also it contains pornography advertisements that diverts the children' mental status. Also, sometimes it may have some political and religious view that causes the unnecessary violence. According to a survey result, a spam mail consists of 50-60% of total incoming emails in the network.

Keywords:Spam,Ham,Falsenpositive,Falsenegative

I.INTRODUCTION

In the world of globalisation there are 4.3 billion users of email,average number of email received by office worker receives is 121 email per day .On an average world is sending 1,49,513 email each minute .So by simple calculation people can sent 269 billion per day .In future yearly growth of Email is projected as growth of 3% . As of the most recently reported period, spam messages accounted for 59.56 percent of e-mail traffic worldwide [22].Spam is the subset of electronic mail that we don't want in our mailbox and email which is desired is known as Ham .Filters are used to categorise emails in Spam & Ham .The statistics about an email filter's effectiveness are built upon the concept of False positives and False negatives . A normal Spam filter that categorizes mail into a spam and a ham category produces a false positive when it puts a ham message into the spam category and a false negative when it places spam into the ham category[24].False positives and false negatives become vitally important only in relation to spam filtering. A false negative (spam that ends up in your inbox) is annoying. A false positive may be an important message that ends up in your spam folder or,

much worse, gets deleted. People tend to be much less bothered by spam slipping through filters into their mail box (false negatives), than having desired email ("ham") blocked (false positives). Trying to balance false negatives (missed spam) vs. false positives (rejecting good email) is critical for a successful anti-spam system. Some systems let individual users have some control over this balance by setting "spam score" limits, etc. Most techniques have both kinds of serious errors, to varying degrees. So, for example, anti-spam systems may use techniques that have a high false negative rate (miss a lot of spam), in order to reduce the number of false positives (rejecting good email). [23]

1.1 The Definition of Spam

Spam is an internet abuse which is a drawback result of a weak vulnerable communication network. Email spam, also known as Unsolicited Bulk Email (UBE), junk mail, or Unsolicited Commercial Email (UCE), is the practice of sending unwanted email messages, frequently with commercial content, in large quantities to an indiscriminate set of recipients directly or indirectly having no personal contact with the recipients. Users and administrators are searching for more and more efficient way to differentiate the spam mail from genuine mails.SPAM is the subset of electronic mail (internet slang) that we do not want in our mailbox. The shortest definition among all the descriptions of spam email is termed as " Unsolicited Bulk Email" [3]. These junk mails are termed as internet slang that are usually the identical mails sent to numerous recipients. In this the sender do not have any personal contact with the receiver.

1.2 History of Spam

Spam in email started to become a problem when the Internet was opened up to the general public in the mid-1990s. The origin year of spam mail was near to 1978.The starting period of spam mail was from 1978 to 1990. In that spam era, the spam were sent manually. As the spammers needed a huge amount of human resources that time, that's why it was not easy to send a much more amount of spam mail to many more users. In that years the spammers send the spams one by one to the users so at that time period email users did not suffer more in spam matters. About 1994, the spam king Je Slaton introduced a spamming program which was able to send millions of spam at a time. This makes the spam as a business tool. About 1995 it became a serious problem when a spam email sent to 2 million addresses at the same time and by the same sender/spammer. As a solution the first spam filter was introduced in 1997 which was the rule based filter as in 1997 when spam showed an exponential growth. As due to the filter was rule based, within a little time delay the spammers get able to break the filter's boundaries by making some smart contents of the spam and hence the filter was no more useful for filtering the spam. The third period starts with the starting of 2002 when Paul Graham published ' A Plan For Spam' which introduces the machine learning and statistical approach of spam mail filtering with a focus on Bayesian networks. He wrote ," I think it's possible to stop spam, and that content-based alters are the way to do it". The Bayesian classification provided the first more efficient spam filter. From that time everyday a new and better spam filter is being modeled to dominate the spammers' activities.

1.3From where the spammers collect e-mail addresses to send spam: The answer might be from chatrooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. Spammers usually don't put much effort into verifying email addresses, they use automatic programs called bots to scour the web and Usenet newsgroups, collecting addresses, or buy them in bulk from other companies. Spammers get the the mass of email addresses from the different search engines. Spammers also guess at addresses using name generation programs, and even send thousands of messages that bounce. As we easily provide our email addresses to the e-shopping organizations and hence their websites such as Flipcart.com , Myntra.com etc, the spammers retreive a bulk of email addresses to send spam from these shopping websites

1.4 Spam contents: A spam may be of different forms as image spam, blank spam, sms spam, email spam etc. The spam mail usually contains advertisement contents. These spam mails in general contains the forms of contents as Phishing scams (email fraud), Foreign bank scams or advance fee fraud schemes, multilevel marketing, accosting by absurd views as "Get Rich Quick" or "Make Money Fast" schemes, fast relief health products and pornographic web sites, Owners of free software's, Chain letters, Illegally pirated software etc.Spam is basically used as a purpose of Advertisement ,Pyramid schemes (Multi Level Marketing) ,Chain Letters ,Political email ,Stock market advice .

1.5 Harm from spam emails : Spam is a cheapest and illegal way to advertise and promote sales and activities as millions of users are reachable on internet . Existence of spam mail causes the unnecessary network traffic, consumes computer resources and time, wastage of memory, bandwidth loss, slow down of mail server, cost shifting and identity theft etc. Finally it causes direct or indirect financial loss or other types of damage. Some spam is annoying but harmless, but some might be part of an identity theft scam or other kind of fraud. According to the estimation of Ferris research analyzer in 2005, the financial loss due to spam were 50 million dollar[2]. Cost of productivity loss from inspecting and deleting spam missed by spam control products (False Negatives) , cost of Productivity loss from searching for legitimate mails deleted in error by spam control products (False Positives), and Operations and helpdesk costs (Filters and Firewalls instalment and maintenance), these causes the direct financial harm due to spam. It sometimes contains harmful and malicious virus that get activated after one click at its link and corrupt our system. Also it contains adult advertisements that divert the children mental status. Also, sometimes it may have some political and religious view that causes the unnecessary violence. Spam is a way used by an individual or any organization to convey their message via advertisement among the public to promote their events and commodities by avoiding the pro and cons of that message. According to a survey result, a spam mail consists of about 70% of total incoming emails in the network [3]. According to google search engine, the statistics of 2005 shows that 67% of the Internet users dislike spam and 33% of them not only dislike spam, but also feels them frustrating, while only 33% of the users have no problem with it.Further, spam accounts for an estimated 70% of global emails – which comes out to approximately 14.5 billion emails per day (source: spamlaws.com).According to Microsoft's Justin Rao and Google's David Reiley (who previous worked together at Yahoo), spam costs the global economy close to $20

billion every year in lost productivity while spammers as a whole make about $200 million per year. How much does spam cost the world? Businesses are falling victim to harmful email and software spam which could set them back thousands of Dollars each year. 94 billion spam threats are sent each day throughout the world, causing detrimental effects to small businesses who struggle with the consequences. In terms of a worldwide figure, $1.6 billion is lost in productivity annually due to unsolicited emails. For example, the most targeted industries included energy and manufacturing industries, spending 6% more on IT to fix these issues, with a 2.8% loss of revenue; financial services at a 7% spend and 1.5% loss, and finally real estate with a 3% spend and a 0.8% loss.According to a survey carried out in 2014, 68% of businesses suffer serious spam threats at least once each year, whilst 45% suffer at least three per year. [22].Businesses can lose money through spam in a variety of ways, such as: Lower employee productivity ,Disinfecting computers ,Losing customers , Lost bandwidth

1.6 What should one do to avoid the spam : As spam causes a lot of direct and indirect loss to the users..Some of anti-spam based practices are listed here:Turn off email preview ,Do not use auto responder Do not read spam ,Do not click on URLs in spam Use a good spam filter ,Give your email addresses only to closely trusted acquaintances.

1.7 Aim towards the spam problem : Filtering spam is the technique to categorise all the incoming emails in network into spam and ham messages. This is not possible to maintain 100% accuracy and efficiency of filtering spam. But one should try to make sure that the model is as more efficient, reliable and accurate as possible.. As a good classifier or filter should avoid the following two schemes:

Ham misclassification: The genuine mail should not be classified as a spam mail. Due to this misclassification the receiver may get unaware of important mails which may be very damaging sometimes by causing serious risks.

Spam misclassification: The spam should not be classified as important mails as it causes many more financial and behavioural damage.

1.8 Process of spam filtering : On that basis the spam filters are of 3 types on the strategy of focusing on emails to classify spam: **(i)** Subject of message

(ii) Body content of message (content based filter)

(iii) Senders status (sender's reputation based on past history as spammer or not)

A general machine learning based spam filter consists of at least the following sequences:

Collection of emails - Firstly all the network emails are collected from individual users which are considered as both spam and legitimate email.

Pre-processing- The next is the transformation process. In this phase the task of prepro-cessing is usually de ned by the author that what strategies she/he is using. Generally it consists of removal of conjunctions, stop words etc. It also have tokenization process in it.

Feature selection - Here in this phase among all the words and attributes those words are selected which are more informative and participate more in filtering spam. Feature selection occurs according to the authors' choice that which methodology he/she is using.

Machine learning application - Now, among all the different machine learning/data mining algorithms (supervised learning, unsupervised learning), one can use either a single or a hybrid of algorithms for rule generation.

Rule extraction- According to the algorithms used some rules are extracted for the given email that play an important role to classify filter email as ham/spam.

Classification - Finally the network email is classified as spam or ham and whether to send this email to user or not.

As per **figure 1**, the pre-processing is performed after the collection of all network emails. Lastly the machine learning techniques are used to provide training and testing to the mails whether to decide the decision as ham/spam.

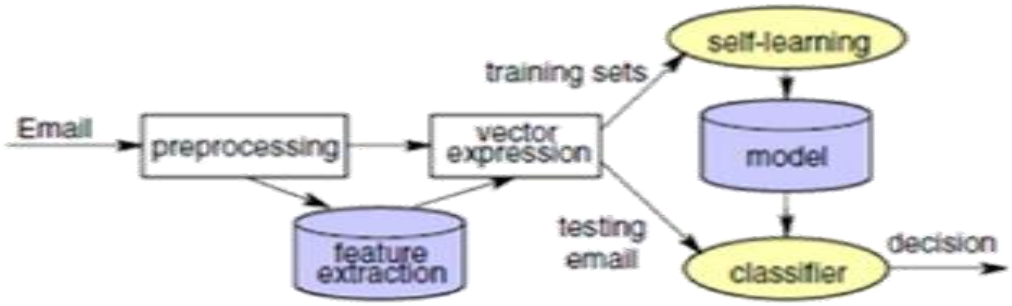

Figure 1: Spam Filtering Process

Now, moving towards the strategy of different levels of filtering spam from the email family Figure 3 shows it's better classification layers. Figure 3 shows the level by level categorization of the network emails ..

II.ANTI-SPAMTECHNIQUES

Spam emails are loss of individual productivity and financial loss of organizations. Antispammers, therefore, are putting forward efforts to prevent this potential threat to today's Internet[3]

Anti-spam techniques can be broken into four broad categories: [2]

Those that require actions by individuals.

1. Those that can be automated by email administrators.
2. Those than can be automated by email senders and
3. Those employed by researchers and law enforcement officials.

In past work has been proposed for a method for classifying e-mails into spam and non-spam. First, several e-mail content features are extracted and then those features are used for classifying each e-mail individually [4]. The classification results of different classifiers (i.e.Bayesian Classification with dynamic training , Decision Trees,A Novel Approach Toward Spam Detection Based on Iterative Patterns ,,K-Nearest Neighbour& others) are used as a mode to study the Spam & Ham difference .Some people ideate with various voting schemes (i.e. majority vote, average probability, product of probabilities, minimum probability and maximum probability) for making the final decision.[18] .E-mail spam has a serious negative impact on the productivity of e-mail using entities in terms of time, money and network resources. Balancing false negatives (i.e. spam e-mails stored in the inbox folder, which annoys the user) and false positives (i.e. good emails transferred to the spam folder, which leads to a loss of valuable information) is considered a critical process to maintain the maximum level of satisfaction of e-mail service subscribers [4,5].

III.LITERATURE SURVEY

This literature survey has not covered all the past and present work performed with the existing algorithm but it tries to cover some of the related work from algorithm of data mining effective in asexample of Spam filtering. This includes Bayesian Classification with dynamic training Decision Trees,Various Decision Trees, K Nearest Neighbour Algorithm,Support Vector Machines, etc with some extra features or with some additional methods in it.

3.1Bayesian Classification with dynamic training

This subsection has covered the Bayes theorem approach with the assumption of the variables within a class are conditional independent of each other [5]. According to the author, unlike other machine learning methods of spam filtering, this method processes words at initial stage instead of simply calculating the word frequency. The paper [5] includes a two step filtering method as shown in figure 2. In figure 2 the classification stage is termed as the testing

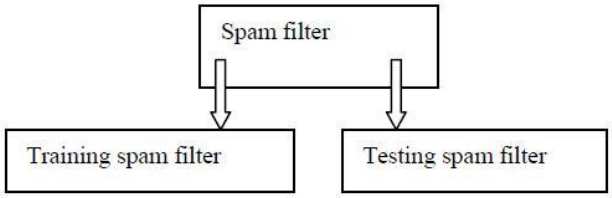

Figure 2: 2 Step Spam Filter

stage[5]. In this 2 stage process the first one is the dynamic training (due to the frequent and regular training and processing of words from the initial stage it is called dynamic) provided to the emails where the word probability (using Bayes theorem) is calculated and the second one is to classify emails as ham/spam. In this paper the Bayes theorem is used to calculate the probability of words such that to decide about the spam probability of mail as a whole. As per Bayes theorem: The Conditional probability of item X belongs to class C i.e. P (C j X) for the known attribute description of X .

P (C j X) = P (C) P (X j C)=P (X):

Now, according to this paper [5], the simple Bayesian algorithm is weak in case of self learning and self adaptability. So, this paper includes the advantage of dynamic training to the Bayes theorem into a single algorithm. From several sources the emails having known label is collected Next to pre-processing phase the useless words like conjunction and stop words are re-moved and a record of sender's status is maintained.Calculate the occurrence frequency of each word and hence maintain a table of entries with their count. On the basis of word probability of having ham/spam the spam probability is calculated as the ratio of frequency of spam to the frequency of combined ham and spam.Now, at last the mails having spam probability more than 0.5 is termed as spam mail.Moving towards another related work to Bayesian classification [6]. To overcome the complicated solutions of Bayes calculation in case of large data sets, a rough set theory is used by the author [6] to reduce the feature set extracted to make the basis of classification. As a new approach that uses public spam database as a new concept [7].

Limitations: One of its weaknesses is the lack in its self learning unlike the decision trees.

3.2Decision Trees : This one is the most popular classification technique for spam filtering. As in [8] some of the decision trees are discussed because of they are easy to implement and understand. The aim is to train the model which provides a target output for several given inputs. Now the question is this, since a simple decision tree is capable enough for classification then why we use additional functions to it. The answer might be that , in simple decision tree the leaf nodes provide a class label which is better for the given data set but it becomes complex for large data sets. Now when we add the extra functions to the child nodes, it gives a function as its leaf node such that we can adjust result accordingly the data sets available [8]

3.2.1Naive Bayes Tree Classifier : This classification approach combines the advantages of naive Bayes classifier with decision tree. In this at each node of decision tree the Bayes rule is applied. This is suitable when database is of arbitrary size and attributes are not necessarily independent.

3.2.2 C 4:5/ J48 Decision Tree :C 4:5 is a well known classification technique. This algorithm chooses some attributes at every node to further classify samples into subsets where each leaf denotes a class or a decision. If all the samples belong to the same class then the tree only gives a leaf node of a single class. J48 is an open source implementation of C4:5. At each node the information gain is calculated and on the basis of more informative attribute, the sample on a node further splits into it's subsets. Figure 3 provides a table of spam

features. This is used at the time of feature selection when the more informative words having high information gain is selected for the purpose of node splitting.

Spam Features	
1	From Correct Domain Name
2	Blocked IP
3	Content Type
4	To header original
5	Is subject present
6	Is reply message
7	Is forwarded message
8	Sensual message
9	Subject content has vulgar words
10	Character set includes foreign language except english

Figure 3: Scheme of Rules assigned to Spam Features

3.2.3 Logistic Model Tree Induction : This model consists of the advantages of decision tree with logistic regression. As regression is about making predictions and is a model of correlation between two or more variables. Now, logistic regression which deals with the binary classification maps a point x in d-dimensional feature space (features can be continuous or categorical) to a value in range 0 to 1.This logistic regression is performed to the tree at each of its leaves. The tree generated by logistic regression model is more accurate and smaller but it takes more time. Logistic regression and decision tree are special case of logistic model tree. As logistic regression is the linear regression function with categorical data sets. The logistic tree model gives in general the 90% of accuracy level to classify the mails.As per figure 4 , it provides an experimental result of the comparative study of the 3 decision trees as logistic model, C4:5/ J48 and Naive Bayes decision tree. Now, moving towards the experimental results, LMT provides the accuracy of 86% and the false positive rate is much lower than other. NB requires highest training time. Here the authors have used the cross validation to predict the accuracy. J48 is considered to be the best whenever training time is being considered as a critical parameter because it takes minimum training time than other DT algorithms discussed here [8].

Limitations of the three: LMT has higher accuracy but it takes high working time, the same happens with NB tree. Now, in case of C4:5/J48, it takes less time to perform but fails in accuracy level.

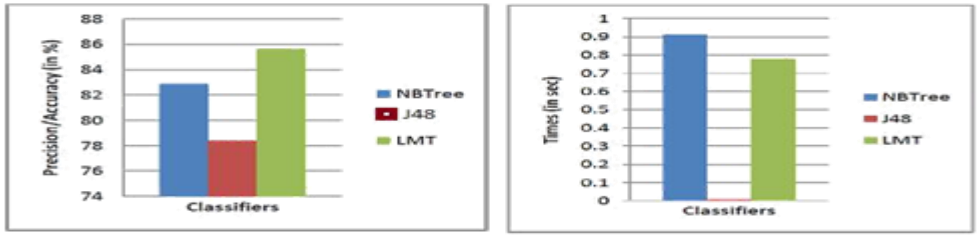

Figure 4: Comparative study of the 3 decision trees

3.3 K Nearest Neighbour Algorithm

Here [9] in case of spam filtering the feature selection is performed first. The attributes like attachments, sender status (reputation based on past records or received mails), message length, links etc are collected and saved and incoming mails are analyzed. Now, on the basis of saved attributes, when a mail needs to be categorized, K most similar documents (K nearest neighbors to that node or mail) in the training set are analyzed. Now, among all the K neighbours if there exists more neighbour having attributes of a spam, then categorize the message as spam else if there exists more neighbours having ham attributes, this will be a genuine mail. Figure 5 shows the sequence of processes proposed in the method of related work done in [9]. Now as per figure7 ,the process takes place as:

- Collection of mails
- Analyze message properties i.e. number of recipients, number of replies, subject length, message size, number of attachments.
- Remove the html tags.
- Feature extraction is performed.
- Remove most used words and update word frequency.

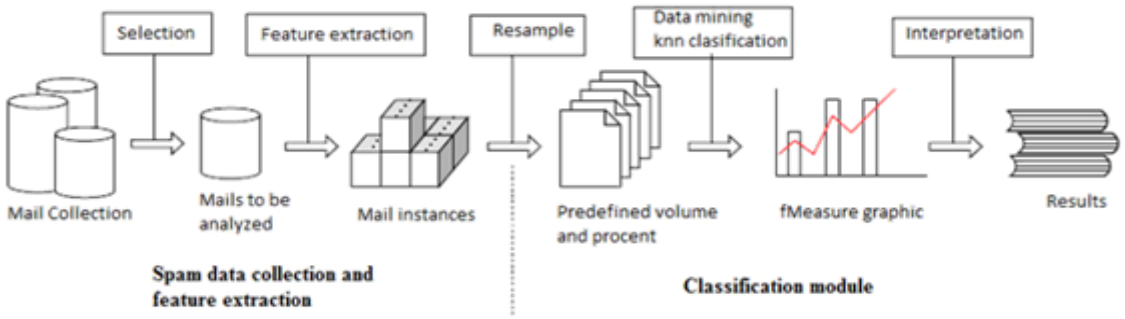

Figure 5: Sequence of proposed Knn algorithm

- Resampling of available data set into optimal size and distribution.
- Classification as spam/ham using Knnalgorithm.

Limitations: It does not perform a feature selection on the extracted attributes. As it causes the unaware behaviour towards the extracted features and hence spam misclassification sometimes.

3.4 A Novel Approach Toward Spam Detection Based on Iterative Patterns

Here in this method [10], some of the iterative patterns (iterative patterns are simply those sequence of words that exists in the sentence more than some threshold and their count of occurrence is called their frequency) are used as a base to classify the incoming mails as spam and ham. Given method is performed on 6 different datasets and it results with higher accuracy. The extracted iterative patterns work as features for the next process of supervised learning. In the previous spam filtering techniques that work on signature based methods were unable to detect the spam over ham when spammers changes regularly the different email addresses and IP addresses. Now next is the rule based techniques. Again it results in high false positive rates. Now the authors[10] have considered the combination of pattern generating technique(frequent item sets generation) with feature selection and then finally the supervised learning. The proposed method consists of 4 processes as:

Pre-processing: This step includes the tokenization, stop words removal and stemming processes.

Term selection: Here TFV(Term Frequency Variance) is used to get best discriminative terms(terms with high variance) because these terms are more informative and others are removed. The TFV is calculated by the given formula [10] as:

$TFV(t_i) = {}^{X}[T\ F\ V(t_i; C)\ T\ F\ V_f^u(t_i)]$

$c(c_s; c_l)$

Where C denotes the class of emails (Cs as spam class and C_j as legitimate email),

$TFV(t_i; C)$denotes the term frequency of ti with respect to class C, and $T\ F\ V_f^u(t_i)$ is the average of term frequencies calculated with respect to both classes.

Feature extraction: The third stage is the feature extraction step by ending the iterative patterns and then selecting the unique and closed patterns which are more informative.

Classification: Lastly, for the classification purpose a classifier is prepared by the learning process from the database tuple database tuples and their associated class labels. Hence finally on that basis the email is classified as spam/ham.

3.5 A Keyword Based Strategy for Spam Topic Discovery from the Internet : Unlike the traditional approaches covers all the non textual information .Spam keywords are the words that cover the spam topics even though having less information. It consists aummary of information related to the spam topic. The methods

used for keyword extraction are dictionary-based approach (good for existing words but not useful in case of new and evolving words), partial meaningful word string recognition (works good for new words but sometimes provide poor theme), multi level filtering candidate spam keyword and multi feature fusion spam keywords weight calculation. Tf-idf factor is used here to calculate the frequency of each term. [12] has provided the efficient formulas for the keyword extraction and hence the spam filtering. The field used here is TDT (Topic detection and Tracking technology) with content based spam filtering through keyword extraction. Authors did their experiments on galaxy public opinion monitoring system of ICT. This provides an efficient spam filtering technique. Limitation: As it is a content based lter so there is a flaw of memory consumption and also that it relies on the end users reporting.

3.6 E-mail Spam Filtering using Support Vector Machines with Selection of Kernel Function Parameters:

Support Vector Machine is a supervised learning model that is used to analyze the data and generate patterns for the purpose of classification and regression analysis. [13] has proposed SVM (Supervised Learning Paradigm) based approach for spam filtering with the inclusion of Taguchi method to improve the grid search methodology of SVM. It uses the orthogonal arrays to avoid iterations. Taguchi method uses 2 tools as Signal-to-Noise ratio (S/N) which measures quality and orthogonal arrays which treats the design parameters. These works as control factors for spam filtering. This method has the unique ability such that it can handle the extremely large feature spaces (such as text). [13] has provided a brief view of SVM theory classification. Here the vector space model (text representing approach) is used to target the emails as a group of orthogonal keywords. This also provides some of the pre-processing to the vector space. Now the Taguchi method is applied. Now, as per the experimental results the accuracy is 97.70%. for the better results the orthogonal table should be enlarged.

Limitation: Proposed approach sometimes may obtain approximation results but not optimal. Accuracy and time taken are inversely proportional here.

3.7 A Novel Method for Image Spam Filtering

Focusing at only the text les is not sufficient for the spam detection. An email consists of text as well as multimedia contents. In this paper the author has provided a view to consider the image spam in terms of spam filtering. As per a recent survey spammer change the whole spam content into image format also one of every 3 spam emails are image spam and this one cant be considered as a formal spam message (textual) and needs special treatment. According to [14] the proposed method is divides into some pieces of tasks as pre-processing of image spam, extraction of connected components (component based method is used to extract the connected components based on some of the dimension and colour properties), classification of connected components(then discard useless components i.e. image or text that do not play a role in identifying spam/ham), merging of components ([14] assumes the Chinese contents so some rules are there for merging the strategies and ow control rules in a single unit), feature selection, use of database to match features and lastly to judge email as spam or ham. The experimental result is conveyed in terms of precision, true positive, true negative, false

positive and false negative. By recognising some special features one can extract image spam. The problem in this paper is how to select the special features when different types of image spam are there.

3.8 Spam Filtering Technique Based on Active Feedback

The idea of having user's feedback towards the network mails to make the classifier more effective and accurate is described in [15]. As, the same mail may be a spam for one person but it may be a ham for another. For example a market sales status may be a genuine mail for some people who want to aware about market trends and sales but for some it acts as a spam or junk mail. So, to solve this problem the author has proposed the idea of users active feedback. Here in this method [15], the mail server sends the mail's information and attribute values of it to the user at a time and asks the user to provide their feedback towards the mail. Now, according to the users feedback the mail server maintain a table of all types of feedback and construct a classifier toward a mail. The mail server maintains a number of classifiers towards every user that get updated time to time according to the user's feedback. Finally, the decision is made by the classifier either to classify that mail as ham/spam for that user.

Limitation : The server has to maintain a number of tables and classifiers, hence it kills time but its performance and filtering accuracy is high as its advantage.

IV.CONCLUSION

Spam is becoming a very serious problem to the Internet community, threatening both the integrity of the networks and the productivity of the users. In this paper, wesurveyed various Data Mining Technique for Anti-spam filtering. In this paper we discussed the problem of spam and given overview of various data mining techniques as spam filtering techniques. There is no common definition of what spam is, but most of the sources agree that the core feature of the phenomenon is that spam messages are unsolicited. Spam causes a number of problems of both economical and ethical nature, which results in particular in the attempts of legislative definitionand prohibition of spam. Email filtering is the processing of email to organize it according to specified criteria. The accuracy of spam blocking techniques are evaluated on two dimensions: How much spam you successfully filter out, and how little legitimate messages you accidentally delete . This can be achieved by introducing modern techniques like behaviour measures & technological measures .

REFERENCES

[1] UpasnaAttri, Satinderpa "A survey of Performance Evaluation Criteria for Spam E-mail Classifiers" in *International Journal of IT, Engineering and Applied Sciences Research (IJIEASR), Volume 1, No. 1, ISSN: 2319-4413, October 2012.*

[2] Sarah Jane Delany , Mark Buckley and Derek Greene, "SMS spam filtering: Methods and data", in *Expert Systems with Applications, ELSEVIER, www.elsevier.com/locate/eswa, 2012.*

[3] Enrico Blanzieri , Anton Bryl, "A survey of learning-based techniques of email spamfiltering" in *Springer Science+Business Media B.V. 2009.*

[4] Omar Saad , Ashraf Darwish, and Ramadan Faraj, "A survey of machine learning tech-niques for Spam filtering" in *IJCSNS International Journal of Computer Science and Network S 66 ecurity, VOL.12 No.2 , February 2012* .

[5] Arun Rajput, DurgaToshniwal, "Adaptive Spam Filtering based on Bayesian Algorithm", in *International Journal on Advanced Computer Engineering and Communication Tech-nology Vol-1 Issue:1 :ISSN 2278 5140*

[6] Yun Wang, Zhiqiang Wu, Runxiu Wu, "Spam Filtering System Based on Rough Set and Bayesian Classier".

[7] Chun-Chao-Yeh and Soun-Jan Chiang, "Revisit Bayesian Approaches For Spam De-tection" in *9th international conference for young computer scientists, IEEE DOI 10.1109/ICYCS.2008.434, 2008*.

[8] Sarit Chakraborty, BikromadittyaMondal, "Spam Mail Filtering Technique using Dif-ferent Decision Tree Classiers through Data Mining Approach - A Comparative Perfor-mance Analysis", in *International Journal of Computer Applications (0975 888) Volume 47 No.16, June 2012.*

[9] LoredanaFirte, CameliaLemnaru, RodicaPotolea, "Spam Detection Filter using KNN Algorithm and Resampling" in *Intelligent Computer Communication and Processing (ICCP), 2010 IEEE International Conference*

[10] Mohammad Razmara, BabakAsadi, MasoudNarouei, Mansour Ahmadi, "A Novel Ap-proach Toward Spam Detection Based on Iterative Patterns", in *2012 2nd International eConference on Computer and Knowledge Engineering (ICCKE), October 18-19, 2012*

[11] Ra qul Islam and Yang Xiang, member IEEE, "Email Classification Using Data Reduction Method"

[12] YongqinQiu , Yan Xu, Dan Li, Hengxun Li, "A keyword based strategy for spam topic discovery from the Internet", in *2010 Fourth International Conference on Genetic and Evolutionary Computing, IEEE DOI 10.1109/ICGEC.2010.71*

[13] Hsu Wei-Chih, Tsan-Ying Yu, "E-mail Spam Filtering Using Support Vector Machines with Selection of Kernel Function Parameters", in *2009 Fourth International Conference on Innovative Computing, Information and Control, IEEE, 2009.*

[14]Hailing Huang, WeiqiangGuo, Yu Zhang, "A Novel Method for Image Spam Filtering", in *The 9th International Conference for Young Computer Scientists, 2008 IEEE DOI 10.1109/ICYCS.2008.440.*

[15]ShaohongZhong, Huajun Huang and Lili Pan, "An E ective Spam Filtering Technique Based on Active Feedback and Maximum Entropy", in *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), IEEE Circuits and Systems Society, 2010.*

[16] Yun-Ju Cho, Hyeseon Lee and Chi-Hyuck Jun, "Optimization of Decision Tree for Clas-si cation Using a Particle Swarm", in *IEMS Vol. 10, No. 4, pp. 272-278, December 2011.*

[17] Bashar Al-Shboul, Heba Hakh1, HossamFaris, Ibrahim Aljarah& Hamad Alsawalqah , "Voting-based Classification for E-mail Spam Detection" ,*J. ICT Res. Appl., Vol. 10, No. 1, 2016, 29-42*

[18] Thiago S. Guzella *, Walmir M. Caminhas ," *A Review of Machine Learning Approaches to Spam Filtering*, Expert Systems with Applications, 36(7), pp. 10206-10222, 2009.

Books

[19]Jiawei Han ,MichelineKamber and Jian Pei, "Data Mining: Concepts and Techniques".

[20] Margaret H Dunham, "Data Mining: Introductory And Advanced Topics".

Internet References

[21]https://www.statista.com/statistics/420391/spam-email-traffic-share/

[22] http://www.radicati.com/wp/wp-content/uploads/2017/01/Email-Statistics-Report-2017-2021-Executive-Summary.pdf

[23] Rushdi Shams and Robert E. Mercer ,"Classifying Spam Emails using Text and Readability Features" ,Email: rshams@csd.uwo.ca, mercer@csd.uwo.ca