IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Email Spam Classification using Hybridized Technique with Feature Selection

Gurwinder Kaur[1], Rupinder Kaur Gurm[2]

Research Scholar, Department of CSE, RIMT-IET, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India[1]

Assistant Professor, Department of CSE, RIMT-IET, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India[2]

Abstract: Email has becomes the major source of communication these days. Majority of people are using this mode of communication for their personal or professional use. Email is an effective, faster, secure and cheaper way of communication. The importance and usage for the email is growing day by day. It provides a way to easily transfer information globally with the help of internet. Because of extensive use of emails, spamming is growing day by day. According to the investigation, it is reported that a user receives more spam or irrelevant mails than ham or relevant mails. Spam is an unwanted, junk, unsolicited bulk message which is used to spreading virus, Trojans, malicious code, advertisement or to gain profit on negligible cost. Spam is a major problem that attacks the existence of electronic mails. So, it is very important to distinguish ham emails from spam emails, many methods have been proposed for classification of email as spam or ham emails. Spam filters are the programs which detect unwanted, unsolicited, junk emails such as spam emails, and prevent them to getting to the users inbox. Machine learning techniques, such as Naïve Bayes, Support Vector Machine, Bagging and decision tree etc. In this paper we introduce a new Hybrid Technique with bagging to enhance the accuracy and performance of classification of emails into spam and ham.

Keywords: Ham, Spam, Email Spamming, Spam Filter, Email Spam.

I. INTRODUCTION

Email becomes the greater source of communication nowadays. Most people on the earth use email for their professional or personal use. Email is a compelling, cheaper and speedier way of communication. It is normal that the aggregate number of worldwide email accounts is expended from 3.3 billion email accounts in 2012 to more than 4.3 billion before the end of year 2016 [email statistic report 2012] . These days, almost every next person in the universe has an email account. The significance and use for the email is increasing every day. It gives a path to easily exchange information universally with the help of web.

In the developing time of web, there is rapid increment in the number of email users which have resulted in the expansion of Spam Emails in late couple of years. Spams are undesirable mails which are sent in enormous amount to anyone anyplace and are of no utilization to the beneficiary. Text classification or text identification represents email classification presents various challenges because of the huge piles of the documents. In numerous datasets, a little percentage of useful features are great in classifying documents and considering all the features can influence the performance. The different mail classification algorithms are SVM, Naive Base, Neural Network, Adaptive boosting algorithm and J48.

Spam is an unwanted, junk, unsolicited bulk message which is circulated to spread virus, malicious code, Trojans, advertisement or for profit on negligible cost. Spams are of various types based on the path by using which it can be transmitted i.e. email spam, web spam, social networking spam, text message spam, blog or review platform spam, instant message spam and comment spam. Spam message can have text, image, video and also voice data. Spam can be sent by means of web, fax, telephonic, sms (text messages). Email spam is increasing day by day in view of growing up of email usage.

The email spamming is expanding day by day due to effective, fast and shabby way of transferring information with each other. As per the investigation, it is reported that a user gets more spam or illegitimate mails than ham or legitimate mails. Around 120 billion of spam mails are sent every day and the cost of sending is roughly zero. As per a spam report of Symantec, the spam rate for December, 2015 was 53.1 percent. Spam not just wastes user time, vitality, consumes resources, storage, bandwidth, computation power but also irritates the user with unwanted messages. Let us take an example, if you received 100 emails in a day. Then roughly about 70 emails are spam and just about 30 emails are ham. In this way, it requires time to distinguish the ham or legitimate emails from it, which bothered the user. Email user gets hundreds of spam emails per day with a new address or email id and new data which are automatically generated by robot programming.

Email is a spam email if it meets the following:

1. Unsolicited email: - The email received by recipient which is not requested by recipient.
2. Bulk mailing/mass mailing: - The email which is sent to large number of people.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

3. Nameless emails: - The email received by recipient in which the address and identity of the sender are hidden.

Spam emails cost billions of dollars every year to the internet service provider in view of the loss of data transmission. Spam emails causes serious problem for internet service provider (ISP), intended user and an entire internet backbone network. One of the case to classify it, might be denial of service where the spammers send mass emails to the server thus delaying legitimate email to reach the intended recipient. Spam is a noteworthy problem that attacks the presence of electronic mails. So, it is essential to recognize ham emails from spam emails, numerous methods have been proposed for classification of email as ham or spam emails.
Spam filters are the programs which distinguish junk, unsolicited and undesirable emails like spam emails and forestall them to getting to the users inbox.

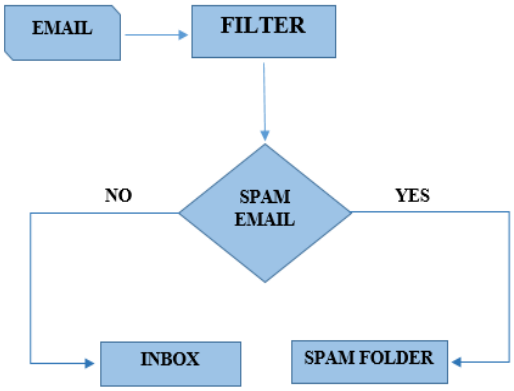

Fig 1. Flow chart of Spam filters

It is found that classification based on non-machine learning success ratio is very low as compared to classification based on machine learning.

Machine learning techniques are neural network, naïve Bayes, support vector machine, and decision tree etc.

Non- Machine learning techniques are signatures, heuristics, black/white list, Mail heading checking etc.

The email is classified into ham or spam by extracting features from an email. Therefore the email classification is based on two feature:

1. Header based features:
Header based Features takes features like sender address, receiver address, subject, bcc, cc etc.

2. Content based features:
Content based Features takes Body regarding feature of the email. Or we can say based on whole of the text which email takes.

Both the set of features to detect spam emails have their own pros and cons. Header features can easily bypassed by the spammers.

II. RELATED WORK

Rushdi Shams and Robert E. Mercer (2013), they reported a novel spam classification technique that uses features, in view of email content language and readability features are combined with the previously used content based task features. The four benchmark datasets such Ling Spam, CSDMC2010, Spam Assassin, and Enron-spam are used to extract the features. They have explained all these features in their paper. These Features are divided three categories such as traditional features, test features, and readability features. The proposed method is capable to classify emails in any language because the features are independent of language. Five well-known machine learning algorithms were used by them to create a spam classifier: Adaboostm 1, Random Forest (RF), Bagging, support vector machine (SVM) and Naïve Bayes (NB). They analyse the classifier performances and they concluded that Bagging performs best out of five listed above. At the end they compare their proposed algorithm to that of many state-to-art anti-spam filters and concluded that their proposed method can be better means to classify spam emails. [1]

Anirudh Harisinghaney and et.al (2014), the main objective of their work is to detect text as well as images as spam emails. For this they tried K- Nearest Neighbor, Naïve Bayes and a newly proposed method Reverse DBSCAN i.e (Density-based spatial clustering of application with noise). They use Enron corpus dataset of text as well as image for experiment purpose. They uses Google's open source library called, Tasseract to extract words from images. Pre-processing of data is performed. They show that with pre-processing all the three algorithms give 50 percent better accuracy results than without using pre-processing. The authors concluded that naïve Bayes algorithm with pre-processing gives the best accuracy among other algorithms listed above. [2]

Masurah Mohamad and Ali Selamat (2015), the authors presented a hybrid feature selection method, known as The Hybrid Feature Selection, in which they combine the rough set theory and TF-IDF to increase the performance result in email spam filters. They explain Feature Selection Methods such as Information Gain (IG), X^{2}-Statistic, Gini Index, Fuzzy Adaptive Particle Swarm Optimization (FAPSO) and Term Frequency Inverse Document Frequency (TF-IDF). And in this case, they provide an explanation of Machine Learning Approaches such as Naïve Bayes and Rough set theory. They utilize spam behaviours and header section which are non-content based keywords. The authors collect dataset comprises of text messages and images. Then they explain their proposed spam filtering framework. In their experimental work the authors show that rough set theory and TF-IDF were have ability to work together in order to generate more accurate and concise results. But when decision tree and TF-IDF combines it gives the best accuracy among others i.e. 89.4% [3]

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Izzat Alsmadi and Ikdam Alhami (2015), in this they use the data set of general statistic about the email from Google report provided for Gmail account user. They classify the dataset based on two methods.

1) Clustering and Classification evaluation
2) Classification based on Word Net class

For classification they use support vector machine and for clustering they use K-Means algorithm. Three SVM models are evaluated such as 1. Top 100 words-VS-email before removing stop words, 2. Top 100 words-VS- email after removing stop words, 3. NGram terms-VS-email. They also concluded that the True Positive(TP) rate is shown to be very large in each case but the False Positive (FP) rate is shown to be best in case of NGram based clustering and classification .[4]

Savita Pundalik Teli and Santosh Kumar Biradar (2014), in there paper, the author compares three classification techniques such as KNN, Support Vector Machine and Naïve Bayes. She shows that Naïve Bayes gives maximum accuracy among other algorithms that is 94.2%. The author then proposed a method to enhance the efficiency of Naïve Bayes. The proposed method is divided into three phases. In first, the user generates rule for classification, secondly trains the classifier with training set by separating the tokens, and in third, based on maximum token matches, the email is classified as ham or spam. They concluded that the accuracy of classifier algorithm is dependent on properly training the classifier in training phase. The efficiency of Naïve Bayes is enhanced a lot by this change in Algorithm. [5]

Ms.D.Karthika Renuka and et.al (2011), in this paper, the authors compare three classification algorithms such as Naïve Bayes, J48 and Multilayer perceptron (MLP) classifier. They find that MLP accuracy rate is greater among others but takes maximum time to classify the emails. But Naïve Bayes algorithm takes least time that is 0.02 but its accuracy is least. They use filtered Bayesian Learning algorithm with Naïve Bayes to increase the performance of Naïve Bayes. The FBL is used for feature selection. After using FBL the accuracy of Naïve Bayes increases to 91%. [6]

My Chau Tu and et.al (2009), in their research they have used WEKA data mining tool and have applied three algorithms to perform classification task to find the heart disease of a patient which are C4.5 algorithm, bagging with C4.5 and bagging with Naïve Bayes. They have used a 10-fold cross validation/verification to calculate the confusion matrix of each model and then analysed its performance by using precision, recall, ROC space, and Fmeasure. They have concluded that bagging algorithms, basically the bagging with Naïve Bayes, performance and output is the best in their research. They believe that their results will make clinical application more available which might further provide great help in healing CAD. Future improvement strategy is to: firstly they believe that bagging with decision tree and bagging with Naïve Bayes which is quite easy to implement and can be used with some more options which can prompt higher results. Secondly, since bagging approach prompt models that are difficult to analyse so they aim at developing a better bagging modelling technique. [7]

III.FEATURE SELECTION TECHNIQUES

Feature selection technique which is used to overpower the task of converting high dimensional data into its smaller possible parts. Feature selection is considered as the most vital part of text mining and data mining.

- **Gini Index**

It is a non-purity split strategy which was enhanced by induction of decision tree. This strategy considers feature containing the minimal class of information in every message. Greater the estimation of purity, better the feature is.

- **Information Gain (IG)**

It is utilized to quantify the amount in bits of information which can be given to the classification system for the prediction of class. Higher estimation of Information Gain (IG) builds its significance.

- **X^2-Statistic**

This technique is additionally called as the Chi-square test, which is utilized in mathematical statistics to test the independence of two variables or attributes. If $X^2(f_i,c_j)=0$, feature f_i, and class c_j are independent, feature f_i does not contain any information of category. However, higher estimation of $X^2(f_i,c_j)$ shows more category information given by feature f_i.

- **Term Frequency Inverse Document Frequency (TF-IDF)**

TF-IDF is derived from arithmetic branch of mathematics, as a numerical statistic method.

$$\text{TF-IDF}(t) = \text{TF}(t) * \text{IDF}(t)$$

It recognizes the frequency of words in a document by measuring the estimation of relevant words through an inverse ratio of the word's frequency in a document to the percentage of documents, the words appears in them. TF-IDF returns high estimation of percentage (i.e output) if the words are common or similar in a single document or in a small clubbing of documents.

- **Fuzzy Adaptive Particle Swarm Optimization (FAPSO)**

FAPSO is categorized into three phases, which are core feature, selection of subset and spam filtering. The motive of this methodology is to find an optimal feature subset.

IV.SPAM DETECTION MACHINE LEARNING TECHNIQUES

1) J48: - The data mining tool, WEKA has implementation of C4.5 algorithm as J48. J48 develops pruned decision trees. J48 algorithm is an evolution of ID3 algorithm. ID3 algorithm works only with nominal attributes while on the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

contrary J48 works not only on nominal but also on numeric attributes. J48 mainly follows the concept of entropy as like ID3 algorithm. For classification, the decision trees here are generated by J48 can be utilized, therefore it is often referred to as a statistical classifier. J48 will create a decision tree which will describe the conditions for an email to be spam. Then using this logic, the detection of spam email can be done. J48 develops pruned decision trees to decrease the complexity.

2) Bagging: - Bagging is an ensemble machine learning meta-algorithm. The bagging technique increases the prediction efficiency of classifiers. Variance and over fitting is reduced by it. At first the process starts with designing bootstrap samples from available overall training datasets and then they create the bagged predictor. The preciseness and efficiency of machine learning algorithms used in statistical classification and regression are enhanced by use of this algorithm. Decision tree methods are usually used with this algorithm. Bootstrap samples are the new training sets which bagging produces from existing training set. Let us assume a standard training set, say S of size x. Bagging creates k new training sets S_i, each of size x, by sampling from S. For email classification, where m is number of models, those are fitted using the above x bootstrap samples and combined them by voting.

3) Naïve Bayes Multinomial: - Naive Bayes classifier is a machine learning algorithm that is based on Bayes' theorem of conditioned probability. It is an algorithm that is used to recognize an email to be spam or ham. Conditioned Probability is given as

$$P(H/X) = P(X/H)\, P(H) / (P(X)).$$

Where H denotes hypothesis, X is some evidences, P (H/X) is the probability of given evidence (X) holds by the hypothesis (H). P (X/H) is probability of X conditioned on (H) hypothesis. P (H) is prior probability of H, independent on evidence (X). Particularly significant words are there, which are used in spam emails and ham emails. These words have probability of occurring in both types of emails. In addition, these filters don't have any idea of these probabilities and how to handle; so we must train the filter to build them up. After training the word, probabilities are used to compute the chances that an email will be considered as spam or ham. Each specific word/keyword or only the most interesting words/keyword results in the email's spam probability. Then, the emails spam probability is calculated for every word in the emails. The email is marked as spam by the filter, if the total probability exceed over certain threshold. In this way, probabilities are used to compute the chances that an email will be considered as spam or ham. The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

V. PROPOSED WORK

The main objective of this proposed work is to enhance the existing machine learning techniques in detecting spam emails, and raise the classification accuracy. It also reduces the variance of prediction and over fitting. We will use TF-IDF as feature selection algorithm. TF-IDF as mention earlier is a good feature selection technique. Then we propose a hybridized technique in which we hybridize two techniques i.e. J48 and Naïve Bayes Multinomial on basis of average of their probabilities with Bagging. J48 with bagging and Naïve Bayes Multinomial with Bagging give good accuracy and performance. Therefore we will combine these two algorithms to enhance the accuracy, precision, Sensitivity and reduces FP rate, FN Rate

A. Objectives

1) To apply pre-processing phase on the emails raw data and convert into formatted dataset using String to word Vector Filter, removing noise and missing values from the raw data. And apply Feature Selection Algorithm i.e. TF-IDF.

2) To apply Bagged Hybrid algorithm i.e. hybridization of J48 and Naïve Bayes Multinomial with Bagging on filtered data for classifying the emails into ham and spam.

3) To apply Bagged J48 i.e. J48 with Bagging, bagged Naïve Bayes multinomial i.e. Naïve Bayes Multinomial with Bagging and J48 algorithm on filtered data for classifying the emails into ham and spam.

4) To compare and analyse the results of proposed technique with the existing on the basis of following parameters: Accuracy Rate (AR), Sensitivity, Precision, False Positive Rate (Fall-out), False Negative Rate (Miss Rate) and test the unlabeled dataset using the proposed classified model.

VI. METHODOLOGY

1) Collection of raw data and then apply filtering techniques to make that raw data into structured format. For doing the classification, Text pre-processing and feature extraction is a preliminary phase. Pre-processing involves 3 steps:

a) **Word parsing and tokenization:** In this phase, each email splits into words of any natural processing language. As email contains block of character which are referred to as token.
b) **Removal of stop words:** Stop words are the words that contain little information so needed to be removed. As by removing them, performance increases. Here, we made a list of around 320 words and created a text file for it. So, at the time of pre-processing we have concluded this stop word so all the words are removed from our dataset.
c) **Stemming:** It is defined as a process to reduce the derived words to their original word stem. For example, "talked", "talking", "talks" as based on the root word "talk".

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

2) Applying TF-IDF as a feature selection algorithm
3) Applying the Decision Tree J48 algorithm on the collected data.
4) Applying an approach that decreases the variance of the prediction using dataset using combinations with repetitions to produce multisets of same size of the dataset as the size of original dataset with randomization and replacement i.e. bagging. For each multi set the learning algorithm J48 is applied to classify the instances and a model is created and a vote related to that model is generated. The average of all the predicted votes is considered to be the result of the classifier.
5) Proposing a new approach in which Naïve Bayes multinomial with Bagging is used.
6) Proposed Bagged Hybrid Algorithm: Proposing an approach that increases the accuracy and decreases the variance of the prediction using dataset using combinations with repetitions to produce multisets of same size of the dataset as the size of original dataset with randomization and replacement which is bagging. For each multi set the learning algorithm i.e. **Hybridized algorithm that contains Naïve Bayes Multinomial algorithm and J48 algorithm being hybridized on the basis of average of their probabilities** is applied to classify the instances and a model is created and a vote related to that model is generated. The average of all the predicted votes is considered to be the result of the classifier. The improved classifier often has significantly greater accuracy than a single classifier derived from D, the original training data. It will not be considerably worse and is more robust to the effects of noisy data. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers. For prediction, it was theoretically proven that a combined predictor will always have improved accuracy over a single predictor derived from D.

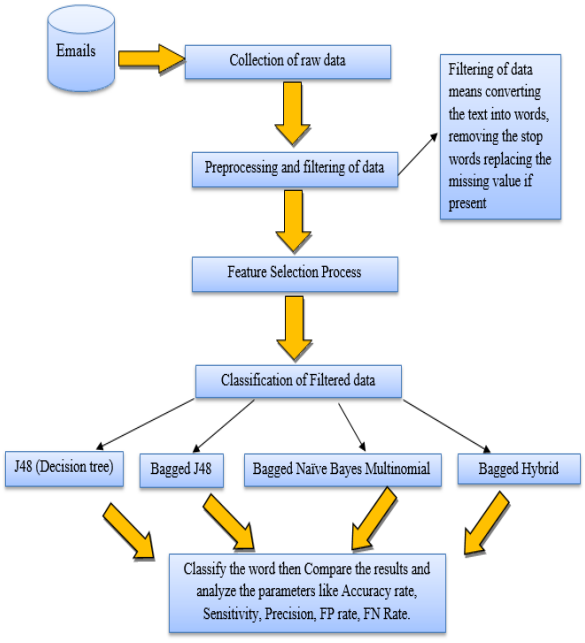

Fig 2. Architecture of Proposed Spam Filtering

5) Analyse the performance parameters like Accuracy Rate (AR), Sensitivity, False Positive Rate (FP Rate), False Negative Rate (FN Rate) and Precision of J48, Bagged J48 ,Bagged Naïve Bayes Multinomial and new Proposed Hybrid algorithms and Compare the results of both and then test for unlabelled data.

VII. EXPERIMENTAL WORK

The work was done to test the performance of the chosen proposed machine learning algorithm, which is the Bagged Hybrid Classification in spam filtering task as a classifier.

This experimental work also tested the ability of a feature selection, which is the Term Frequency Inverse Document Frequency (TF-IDF) which helping the Bagged Hybrid classifier to classify spam messages.

First, the dataset comprising of text emails were collected from our own email inbox and several public data sets. Even though there are many public email spam datasets provided, we preferred to use our own email collection, where the content of the public dataset is approximately similar with our own email collection and to preserve the originality of research work.

Furthermore, the type of data collection is not the major concern in this research as long as the content of an email contains "spam text and ham text". This experimental work had collected training dataset of 201 emails comprising of texts where 110 text instances were categorized as spam, while another 91 text instances were categorized as ham. The test set with unlabelled class contains 25 emails some are spam and some are ham.

Secondly, all of these emails went through a pre-processing phase prior to training and testing processes. The pre-processing phase is also called as the feature extraction process. At this level, all emails were cleaned up in order to remove unnecessary words using the stop word removal and converting strings into words using string to word vector.

Thirdly, after all emails were cleaned up, they went through a feature selection phase. In this phase, a feature selection method was applied, which is the Term Frequency Inverse Document Frequency (TF-IDF) technique. The TF-IDF value was calculated for all words in each document as an input value for classification phase. Fourthly, apply all the four algorithms which are mentioned earlier i.e. J48, Bagged J48, Bagged Naïve Bayes Multinomial and Bagged Hybrid Classification which is the combination of J48 and Naïve Bayes Multinomial.

Confusion Matrix: A table of confusion (sometimes also called a confusion matrix), is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

	True Positive ($n_{s\to s}$)	False Positive ($n_{h\to s}$)	False Negative ($n_{s\to h}$)	True Negative ($n_{h\to h}$)
J48	87	12	23	79
Bagged J48	91	8	19	83
Bagged NBM	105	15	5	76
Bagged Hybrid	105	7	5	84

Accuracy: Accuracy is the percentage of correctly identified spams and hams. It can be measured as the number of correctly classified instances to the total number of instances.

$$\text{Accuracy} = \frac{n_{s\to s} + n_{h\to h}}{n_{s\to s} + n_{s\to h} + n_{h\to s} + n_{h\to h}}$$

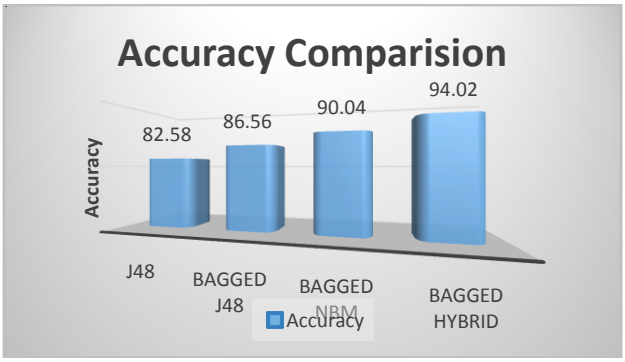

Fig 3. Performance evaluation based on accuracy rate (AR)

This figure shows the accuracy analysis for Bagged Hybrid is around 94.02% which is highest among others, for Bagged Naïve Bayes Multinomial is around 90.04%, for Bagged J48 is around 86.56% and for J48 is 82.58%.

Sensitivity or TP Rate: It checks how many instances are correctly classified a spam. It can be measured by number of instances that are correctly classified as spam to the total number of spam instances.

$$\text{Sensitivity} = \frac{n_{s\to s}}{n_{s\to s} + n_{s\to h}}$$

Precision: It checks how many instances are correctly classified a spam among those all that are classified as spam.

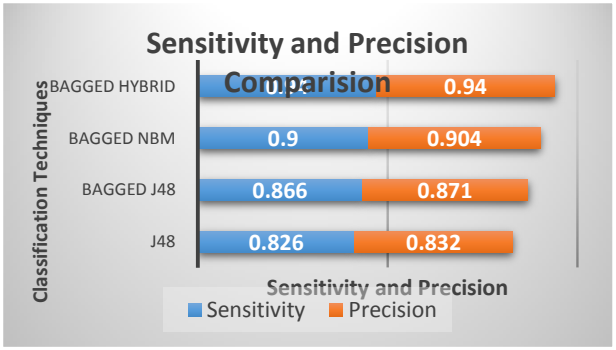

Fig 4. Performance evaluation based on Sensitivity and Precision

It can be measured by number of instances that are correctly classified as spam to the total number of instances classified as spam.

$$\text{Precision} = \frac{n_{s\to s}}{n_{s\to s} + n_{h\to s}}$$

This figure compares the sensitivity, FP Rate and Precision. Shows that Sensitivity and Precision of our proposed method i.e. Bagged Hybrid is highest.

FP Rate (Fall-out): It checks how many instances are incorrectly classified as spam. It can be measured as number of instances that are incorrectly classified as spam to the total number of ham instances. It should be low.

$$\text{FP Rate} = \frac{n_{h\to s}}{n_{h\to s} + n_{h\to h}}$$

FN Rate (Miss-Rate): It checks how many instances are missed to classify as spam which are actually of class spam or we can say how many instances are incorrectly classified as ham. It can be measured as number of instances that are incorrectly classified as ham to the number of actual spam instances. It should also be low.

$$\text{FN Rate} = \frac{n_{s\to h}}{n_{s\to s} + n_{s\to h}}$$

The below figure compares the FP Rate and FN Rate. Which are errors in classification and should be low. Shows that FP Rate for our proposed method i.e. Bagged Hybrid is lowest among others. and FN Rate for our proposed method is same as for the Bagged Naïve Bayes Multinomial but it is very low i.e. 0.045.

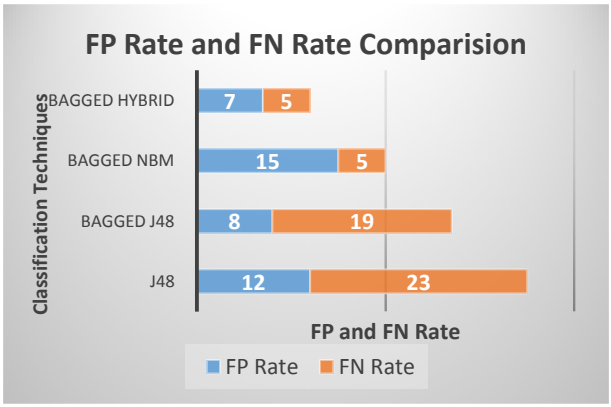

Fig 5. Performance evaluation based on FP Rate and FN Rate

DOI 10.17148/IJARCCE.2016.51259

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

VIII. CONCLUSION

During the work, we have observed that there are numerous email spam detection techniques available around us. These technique either lack in accuracy or level of performance. From all of these techniques no one can reaches to 100% accuracy. The classification depends on content features gives the better results in accuracy than header based. But the accuracy of all these techniques has been enhanced using Feature selection techniques. Therefore feature selections is providing greater role in email spamming. Therefore we are proposing a new hybridize technique i.e. Bagged Hybrid Algorithm. We have concluded that the proposed hybrid technique enhances the accuracy of email spam classification up to 94.02%. We also compare these algorithms on the basis of Sensitivity, Precision, FP Rate, FN Rate and observed that our Proposed Hybrid Algorithm work well to classify emails into spam and ham.

IX. FUTURE SCOPE

I hope this work helps other researchers we can also do this work with boosting rather than Bagging. We can also add other good feature selection algorithms or hybridize feature selection algorithms to enhance more accuracy.

ACKNOWLEDGEMENT

The Author Gurwinder Kaur would like to acknowledge the contribution of **Mrs Rupinder Kaur Gurm** for his valuable suggestions and guidance. Her extremely successful professional life has been a strong motivating factors in pursuit my M.Tech. She is not only a great advisor but also a caring mentor during my M.Tech. Her amazing energy and strong dedication will continue to be the source of inspiration to me.

REFERENCES

[1] Rushdi Shams and Robert E. Mercer, "Classification spam emails using text and readability features," IEEE 13th International Conference on Data Mining, 2013.
[2] Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora , "Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN Algorithm, " International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, India, Feb 6-8 2014.
[3] Masurah Mohamad and Ali Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," IEEE International Conference on Computer Communication, and Control Technology (14CT 2015), April. 2015.
[4] Izzat Alsmadi and Ikdam Alhami, "Clustering and Classification of email contents," Journal of King Saud University-Computer and Information Sciences, vol. 27, pp. 46-57, Jan. 2015.
[5] Ms.D.Karthika Renuka, Dr.T.Hamsapriya, Mr.M.Raja Chakkaravarthi, Ms.P.Lakshmisurya, "Spam Classification based on Supervised Learning using Machine Learning Techniques," IEEE, 2011.
[6] Savita Pundalik Teli and Santosh Kumar Biradar, "Effective Email Classification for Spam and Non-spam," International Journal of Advanced Research in Computer and software Engineering, vol. 4, June 6, 2014.
[7] My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", DASC '09 Proceedings of the 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, IEEE Computer Society Washington, DC, USA ©2009.
[8] Rekha and Sandeep Negi, "A Review on Different Spam Detection Approaches," International Journal of Engineering Trends and Technology (IJETT), Vol. 1, May 6, 2014.
[9] Megha Rathi and Vikas Pareek, "Spam Email Detection through Data Mining-A Comparative Performance Analysis," I.J. Modern Education and Computer Science, vol. 12, pp. 31-39, 2013, available on http://www.mecs-press.org/.
[10] Guanting Tang, Jian Pei, and Wo-Shun Luk,"Email Mining: Tasks, Common Techniques, and Tools", School of Computing Science, Simon Fraser University, Burnaby BC, CANADA.
[11] Gurwinder Kaur and Rupinder Kaur Gurm, "A Survey on Various Classification Techniques in Email Spamming", International Journal of Advance Research in Computer and Communication Engineering (IJARCCE), vol. 5,pp. 589-593 March,2016
[12] Seongwook Youn and Dennis McLeod," A Comparative Study for Email Classification", University of Southern California, Los Angeles, CA 90089 USA.
[13] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, "Comparative Study on Email Spam Classifier using Data Mining Techniques", IAENG.
[14] Geerthik.S," Survey on Internet Spam: Classification and Analysis", Int.J.Computer Technology & Applications, Vol 4 (3), pp. 384-391.
[15] a Sharma, Gurjot Kaur," Spam Detection Techniques: A Review" International Journal of Science and Research (IJSR), 2013.
[16] Enron Spam Public Corpus, http://csmining.org/index.php/enron-spam-datasets.html.
[17] Weka, http://www.cs.waikato.ac.nz/ml/weka/
[18] LingSpam Public Corpus, http://csmining.org/index.php/ling-spam-datasets.html.