INTERNATIONAL JOURNAL OF TECHNOLOGY AND COMPUTING (IJTC)
ISSN-2455-099X,
Volume 2, Issue 3 March 2016.

A Review on Various Classification Techniques in Email Spamming

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

[1]Research Scholar, Department of CSE, RIMT-IET, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India,
[1]er.gurwinderkaurarora@gmail.com
[2]Assistant Professor, Department of CSE, RIMT-IET, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India,
[2]rupindergurm@gmail.com

ABSTRACT

Email becomes the major source of communication these days. Most humans on the earth use email for their personal or professional use. Email is an effective, faster 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. Due to it the email spamming is increasing 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. The filter classification techniques are categorized into two either based on machine learning technique or based on non-machine learning techniques. Machine learning techniques, such as Naïve Bayes, Support Vector Machine, Adaboost and decision tree etc. whereas non- machine learning techniques, such as black/white list, signatures, mail header checking etc. In this paper we review these techniques for classifying emails into spam or ham.

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

I. INTRODUCTION

The data mining is basically "The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions" – Simoudis 1996". Data mining is concerned with the analysis of data for finding hidden and unexpected pattern and relationships in large volume of data. Basically the focus of data mining is to find the information which is

hidden and unexpected and convert it into the understandable form for future use. Data mining is also called as KDD, knowledge discovery in databases.

Data mining techniques are listed below:

1. Classification: Is used to place the data in predetermined group.
2. Clusters: Data items are placed in a group according to logical relationships.
3. Associations: Data mining is applied to data set to find out the associations.
4. Sequential Patterns: Data is mined to expected behaviour patterns and trends.

Knowledge discovery steps are:

1. Data cleaning:- to remove noise , irrelevant and inconsistent data from the database
2. Data integration:- the step where multiple data sources may be combined to build a data set
3. Data selection:- the step where the data relevant to the analysis are selected from the data base
4. Data transformation:- the step where the data are transformed into the form that is appropriate for mining

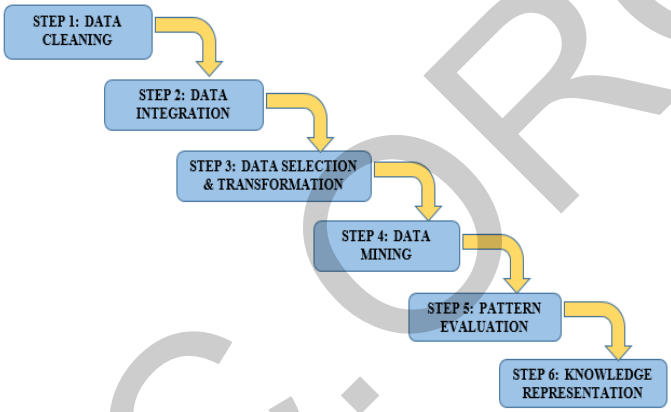

Fig. 1: Knowledge Discovery

5. Data mining: - the process where intelligent methods are applied in order to extract data patterns from data set.
6. Pattern evaluation: - evaluate patterns to identify the truly interesting patterns representing knowledge.
7. Knowledge representation: - where knowledge representation techniques are used to present the minded knowledge.

Email becomes the major source of communication these days. Most humans on the earth use email for their personal or professional use. Email is an effective, faster and cheaper way of communication. It is expected that the total number of worldwide email accounts is increased from 3.3 billion email accounts in 2012 to over 4.3 billion by the end of year 2016[email statistic report 2012] . Now days, almost every second user in the earth has an email account. 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.

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. Spams are of many types based on the way of transmission i.e. email spam, social networking spam, web spam, blog or review platform spam, instant message spam, text message spam and comment spam. Spam message can contain text, image, video and also voice data. Spam can be sent via web, fax, telephonic, SMS (text messages).

The email spamming is increasing day by day because of effective, fast and cheap way of exchanging information with each other. According to the investigation, it is reported that a user receives more spam or irrelevant mails than ham or relevant mails. About 120 billion of spam mails are sent per day and the cost of sending is approximately zero. According to a spam report of Symantec, the spam rate for December, 2015 was 53.1 percent. Spam not only wastes user time, energy, consumes resources, storage, computation power, bandwidth but also irritates the user with unwanted messages. For example, if you received 100 emails today. Then about approximately 70 emails are spam and only about 30 emails are ham. So, it takes time to identify the ham or important emails from it, which irritated the user. Email user receives hundreds of spam emails per day with a new address or identity and new content which are automatically generated by robot software.

Email is a spam email if it meets the following
Criteria:

1. Unsolicited email: - The email which is not requested by recipient.
2. Bulk mailing/mass mailing: - The email which is sent to large group of people.
3. Nameless emails: - The email in which the address and identity of the sender are hidden.

Spam emails cost billions of dollars per year to the internet service provider because of the loss of bandwidth. Spam emails causes serious problem for intended user, internet service provider and an entire internet backbone network. One of the examples to explain it, may be denial of service where the spammers send bulk emails to the server thus delaying relevant email to reach the intended recipient. 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. The filter classification techniques are categorized into two parts:

1. Based on machine learning technique.
2. Based on non-machine learning techniques.

Machine learning techniques, such as naïve Bayes, support vector machine, neural network, and decision tree etc. whereas non- machine learning techniques, such as heuristics, black/white list, signatures, Mail heading checking etc. It is found that classification based on machine learning success ratio is very high as compared to classification based on non-machine learning. The email is classified into spam or ham by extracting features from an email. Therefore the email classifications are based on two feature selection.

1. Header based features
2. Content based features

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

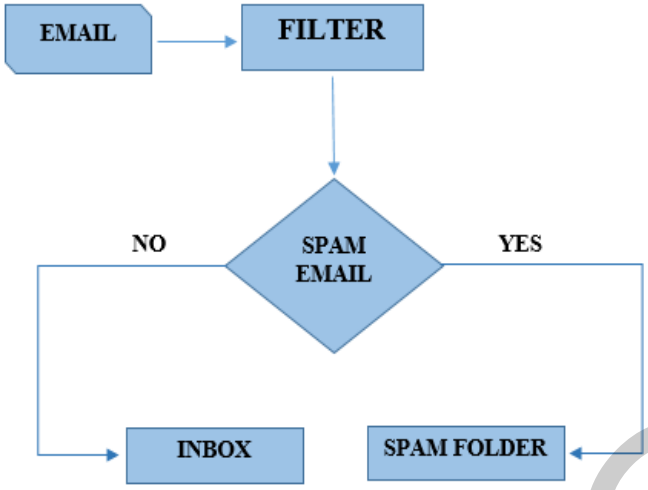

Fig. 2: Flow chart of Spam filters

Outline of this paper:
This paper is organized as follow section 2 presents related work, section 3 presents various feature selection techniques, section 4 comprised of comparison of techniques, section 5 presents proposed work, section 6 comprised of methodology and section 7 presents conclusion.

II. RELATED WORK

Rushdi Shams and Robert E. Mercer (2013) performed a work "Classification spam emails using text and readability features". They reported a novel spam classification method that uses features, based on email content language and readability combined with the previously used content based task features. The features are extracted from four benchmark datasets such as CSDMC2010, Spam Assassin, Ling Spam, and Enron-spam. They explain all these features. Features are divided three categories i.e. traditional features, test features, and readability features. The proposed method is able to classify emails in any language because the features are language independent. They use five well-known machine learning algorithms to introduce spam classifier: Random Forest (RF), Bagging, Adaboostm 1, support vector machine (SVM), Naïve Bayes (NB). They evaluate the classifier performances and concluded that Bagging performs the best out of five. At last they compare their proposed method to that of many state-to-art anti-spam filters and concluded that their proposed method can be a good means to classify spam emails [1].

Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora (2014) performed a work "Text and Image Based Spam Email Classification Using KNN, Naïve Bayes and Reverse DBSCAN Algorithm" The objective of their work is to detect text as well as spam emails. For this purpose they use Naïve Bayes, K- Nearest Neighbor and a new proposed method Reverse DBSCAN (Density-based spatial clustering of application with noise). They use enron cropus dataset of text as well as image. They extract words from image by using Google's open source library called, Tasseract. They use pre-processing of data. They show that pre-processing gives 50 percent better accuracy results with all the three algorithms than without using pre-processing. They concluded that naïve bayes with pre-processing gives the best accuracy among other algorithms [2].

Masurah Mohamad and Ali Selamat (2015) performed a work "An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification". They presented a hybrid feature selection method, namely The Hybrid Feature Selection, in which they integrate the rough set theory and term frequency inverse document frequency (TF-IDF) to increase the efficiency result in email filters. They explain Feature Selection Methods such as Information Gain (IG), Gini Index, X2-Statistic, Fuzzy Adaptive Particle Swarm Optimization (FAPSO) and Term Frequency Inverse Document Frequency (TF-IDF) and Machine Learning Approaches such as Naïve Bayes and Rough set theory. They use header section and spam behaviours which are non-content based keywords. They use dataset comprises of text messages and images. Then they explain their proposed spam filtering framework. In their experimental work they show that rough set theory and TF-IDF were able to work together in order to generate concise and more accurate results. But the combination of decision tree and TF-IDF gives the best accuracy among others i.e. 89.4% [3].

Izzat Alsmadi and Ikdam Alhami (2015) performed a work "Clustering and Classification of Email Contents". In this they explain various research papers based on spam detection, ontology classification on email content and other research goals. 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) Classification based on WordNet class 2) Clustering and Classification evaluation. For clustering they use K-Means algorithm and for classification they use support vector machine. 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 concluded that the True Positive(TP) rate is shown to be very high in each case but the False Positive (FP) rate is shown to be best in case of NGram based clustering and classification [4].

Ms.D.Karthika Renuka, Dr.T.Hamsapriya, Mr.M.Raja Chakkaravarthi, Ms.P.Lakshmisurya (2011) performed a work "Spam Classification based on Supervised Learning using Machine Learning Techniques". In this paper, the authors compare three classification algorithms such as Naïve Bayes, J48 and Multilayer perceptron (MLP) classifier. They evaluate that MLP accuracy rate is higher among others but takes maximum time to classify. And Naïve Bayes takes minimum time that is 0.02 but its accuracy is least. They use filtered Bayesian Learning algorithm with Naïve Bayes to enhance the performance of Naïve Bayes. The FBL is used for feature selection. After using FBL the accuracy rate of Naïve Bayes increases to 91% [5].

Megha Rathi and Vikas Pareek (2013) performed a work "Spam Email Detection through Data Mining-A Comparative Performance Analysis". In this paper the author explains the data mining concept and also the classification algorithms. They evaluate various classification algorithms such as naïve bayes, Bayesian net, random forest, random tree, SVM etc. without feature selection first. Then they evaluates all these classification algorithms with feature selection by best first algorithm. The author evaluated that random tree has 90.43% accuracy, which is very low. But with feature selection it reaches to 99.71% which is very close to 100%. Therefore the author concluded that random tree is the best classification algorithm for email classification with feature selection [6].

Savita Pundalik Teli and Santosh Kumar Biradar (2014) performed a work "Effective Email Classification for Spam and Non-spam" In this 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 phase the user

creates rule for classification, second phase trains the classifier with training set by extracting the tokens, and in third phase based on maximum token matches, the email is classified as spam or ham. They concluded that the accuracy of classifier algorithm is dependent on properly training the classifier in training phase. The performance of Naïve Bayes is improved by this Algorithm[7].

III. FEATURE SELECTION TECHNIQUES

Feature selection techniques are used to overcome the task of extracting high dimensional data into the smallest possible data. Gini Index, Information Gain (IG), X2-Statistic, Term Frequency Inverse Document Frequency (TF-IDF) and Fuzzy Adaptive Particle Swarm Optimization (FAPSO) are among the popular techniques used in spam filtering task.

1. **Gini Index: -** It is a non-purity split method which was improved from a decision tree induction [5]. This method considers feature containing the least category of information in every message. Higher the value of purity, better the feature is.
2. **Information Gain (IG):** -It is used to measure the amount in bits of information which can be provided to the classification system for the prediction of class. Higher value of Information Gain (IG) increases its significance.
3. **X2-Statistic: -** This method is also called as the Chi-square test, which is used to test the independence of two variables or attributes in mathematical statistics. It is applied in the feature selection method in order to determine the independence of a feature fi, and a class cj. If X2(fi,cj)=0, feature fi, and class cj are independent, feature fi does not contain any information of category. Otherwise, higher value of X2(fi,cj) indicates more category information given by feature fi.
4. **Term Frequency Inverse Document Frequency (TF-IDF):** -TF-IDF is a numerical statistic technique from mathematics. TF-IDF is the product of two numerical statistics i.e. term frequency and inverse document frequency. It identifies the frequency of words in a document by measuring the value 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 values of percentage if the words are common in a single document or a small group of documents.
5. **Fuzzy Adaptive Particle Swarm Optimization (FAPSO): -** FAPSO is divided into three stages, which are core feature subset selection, feature subset selection and spam filtering. The objective of this technique is to detect an optimal feature subset.

IV. SPAM DETECTION TECHNIQUES

There are various spam detection techniques. Out of which some are machine learning and some are non-machine learning. Some of them are defined below:

A. Non-Machine Learning Techniques

1) **Blacklist\Whitelist: -** This technique simply creates two lists. A whitelist is a list which includes the email addresses or entire domains which are known to the user. An automatic white list management tool is also use by user that helps in automatically adding email addresses to the whitelist that are known to user. A blacklist where we add addresses that are ambiguous, unsolicited ad harmful for users.

2) **Signatures: -** This technique is based on generating a signature with unique hash value for every spam message. The filters compare the previous stored values with incoming emails values. It is approximately impossible for relevant message to have same value with spam message value that is stored earlier.

3) **Mail Header Checking: -** In this technique we simply create set of rules that we match with mail headers. If a mail header matches, then it triggers the server and return mails that have

- empty "From" field
- Different addresses in "To" field from same source address.
- Too many digits in address etc.

B. Machine Learning Techniques

1) **AdaBoost Classifier: -** Stand for Adaptive boosting, is a machine learning algorithm proposed by Freund and Robert Schapiro. It is a Meta algorithm which can be used in aggregation with some other learning algorithms to improve the performance of AdaBoost algorithm. AdaBoost classifier uses Confidence based label sampling that works with the concept of active learning. Classifier is trained by the variance and obtains a scoring function which is used to classify the mail as spam or ham. The labelled data is used to train the data. The trained classifier generated the required functions which classify the message as spam. This algorithm improves training process.

2) **Naïve Bayes: -** A machine learning algorithm, Naive Bayes classifier is based on Bayes' theorem of conditioned probability. It 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. P (H) – prior probability of H, independent on X. There are particularly significant words used in spam emails and ham emails. These words have probability of occurring in both emails. In advance the filters don't know these probabilities; we must train the filter to build them up. After training the word probabilities are used to compute the probability that an email have that belong to either spam or ham emails. Each particular word or only the most interesting words contribute to email's spam probability. Then, the emails spam probability is computed for every word in the emails. If this total probability exceed over certain threshold then the filters will mark that emails as spam.

3) **Bagging: -** Bagging, also called Bootstrap aggregating, is an ensemble machine learning meta-algorithm. The bagging technique increases the efficiency of the prediction of classifiers. Initially the process starts with designing bootstrap samples from available training datasets and then they produce the bagged predictor. It is used to improve the accuracy and stability of machine learning algorithms used in statistical classification and regression. It is usually applied to decision tree methods. Bagging generates new training sets from existing training set known as bootstrap samples. Assume a standard training set, say T of size n. Bagging creates k new training sets Ti, each of size n, by sampling from T. For spam detection, m models are fitted using the above m bootstrap samples and combined them by voting. It reduces variance and helps to avoid over fitting.

4) **Support Vector Machine: -** In machine learning, support vector machines (SVMs) and also called support vector networks. SVM is a supervised machine learning models that analyse data and make out patterns that are used for classification analysis. Given a set of training examples, each marked as belonging to one of two categories of class attribute. An SVM algorithm builds a model that assigns new examples into one of the two category. That make it a non-probabilistic binary linear classifier. In email spam detection, emails are divided into two classes i.e. "spam" and "ham" by a hyper plane. The aim is to find a hyper plane, which can maximize the margin between the spam and ham classes, this is known as the optimum separating hyper plane. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based

on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high dimensional feature spaces. In email spam classification it gives best result in case of Header based classification.

5) **J48: - J48 algorithm** is basically a java implementation of C4.5 algorithm. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. It is used in the data mining tool, WEKA. J48 creates pruned decision trees. J48 is an evolution of ID3 algorithm. ID3 algorithm works only on nominal attributes whereas J48 works not only on nominal but also on numeric attributes. Like ID3 algorithm, J48 works on the concept of entropy. J48 makes pruned decision trees. The decision trees generated by J48 can be used for classification; therefore J48 is often referred to as a statistical classifier. J48 will create a decision tree which will describe the conditions for a mail to be spam. Then using this logic, detection of spam email can be done.

TABLE 1: COMPARISON OF DIFFERENT SPAM DETECTION TECHNIQUES

Technique	Advantage	Disadvantage
Non-Machine Learning		
Blacklist/whitelist	Simple	Spammers can easily penetrate it.
Signature	Less value of False Positive (FP).	Unable to detect spam until email is reported as spam and its hash is distributed.
Mail Header Checking	Easy to implement.	False Positive (FP) rate is high and requires extra information.
Machine Learning		
Adaboost	Less susceptible to training data overfit.	Suboptimal solutions.
Naïve Bayes	Takes minimum time to detect spam in emails.	Based on ‘naive’ Bayesian filtering, which assumes events are occurred mutually exclusively.
Bagging	Uses a specific or major attribute and perform classification by voting.	Difficult to imply combining rules other than voting.
Support Vector Machine	Dispersion of errors is better.	Takes greater time to classify.
J48	Creates pruned decision trees to decrease the complexity.	Lower Confidence Factor

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 proposing a hybridized technique that improves the efficiency of J48 by using Bagging with it. Because J48 is the best classifier among other decision trees. And bagging

results best ensemble Meta learning algorithm. Therefore we will combine these two algorithms to enhance the accuracy, precision, recall, and TP rate and reduces FT rate.

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 preprocessing and feature extraction is a preliminary phase. Preprocessing involves 3 steps:

a) Word parsing and tokenization: In this phase, each user review splits into words of any natural processing language. As movie review 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 preprocessing we have concluded this stop word so all the words are removed from our dataset i.e. filtered.

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". We have used Snowball stemmer to reduce the derived word to their origin.

2) Applying TF-IDF as a feature selection algorithm.

3) Applying the Decision Tree J48 algorithm on the collected data.

4) Proposing a new approach that decreases the variance of the prediction using dataset using combinations with repetitions to produce multisets of same size of the 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) Analyse the performance parameters like FP rate, TP rate, Recall, Precision and accuracy of J48 and new proposed hybridized algorithm and Compare the results of both and then test for unlabelled data.

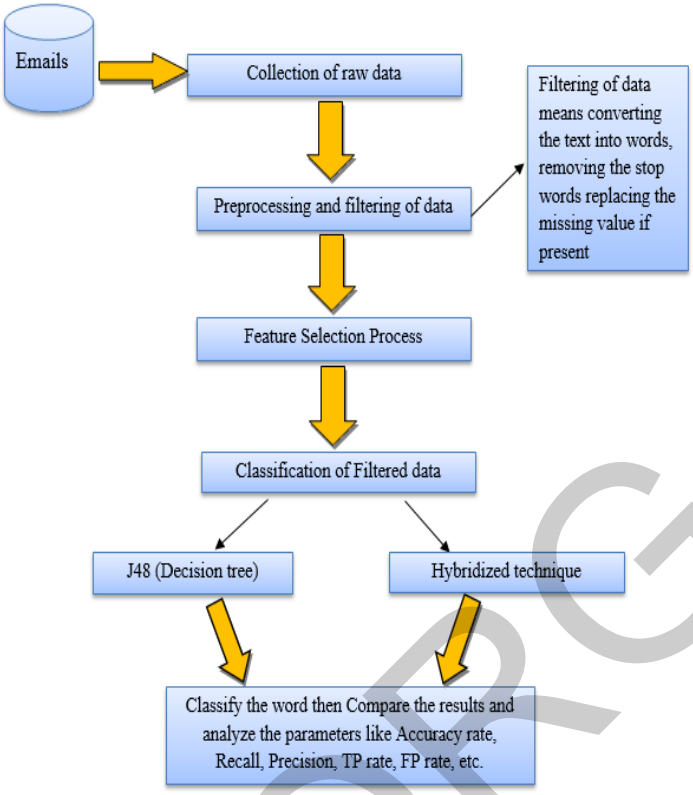

Fig. 3: Architecture of Proposed Spam Filtering.

VII. CONCLUSION

During the survey, we read many research papers and 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. Hope that this hybridize technique enhances the accuracy, precision, TP rate and recall.

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 PG. She is not only a great advisor but also a caring mentor during my PG. 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] 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/.
[7] 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.
[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] 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.
[10] Seongwook Youn and Dennis McLeod," A Comparative Study for Email Classification", University of Southern California, Los Angeles, CA 90089 USA.
[11] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, "Comparative Study on Email Spam Classifier using Data Mining Techniques", IAENG.
[12] Geerthik.S," Survey on Internet Spam: Classification and Analysis", Int.J.Computer Technology & Applications,Vol 4 (3), pp. 384-391.
[13] Reena Sharma, Gurjot Kaur," Spam Detection Techniques: A Review" International Journal of Science and Research (IJSR), 2013.