International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 9 Sep. 2022, pp: 1436-1443 www.ijaem.net ISSN: 2395-5252

Email Spam Detection using Supervised Machine Learning Techniques

Hema Sundar Tatipudi
School of Computer Science and Engineering
Lovely Professional University
Phagwara, India

Date of Submission: 20-09-2022 Date of Acceptance: 30-09-2022

ABSTRACT – Over the past few decades, Technology has gained a rapid pace in its development making communication more easier. Considering several modes of communication, E-mails(Electronic mails) are the best means for both informal and formal conversation. Some also use e-mails to store and share important information in the form of text, images, documents, etc. between people using electronic devices. Besides, some people improperly use this means of communication by sending useless or unwanted e-mails in bulk which could result in disproportionate usage of memory in the mailbox. Such kind of the e-mail messages are considered as Spam.

There are many suggested approaches that could identify spam emails from the mail box. Identifying spam emails mostly involves methods using machine learning whether it may be supervised, unsupervised or reinforcement. Taking several parameters like Accuracy, Error, Evaluation time, Efficiency and so on into consideration; every technique has its own advantages and disadvantages. This review draws the contrast on strengths, drawbacks and limitations of some of the existing techniques that uses the approach of supervised machine learning to detect spam emails.Machine learning method is further resourceful than acquaintanceapproach of engineering; which does not involve the specifications of any instructions. As an alternative, a collection of pre classified e-mail messages is used, these models are a set of training models. To learn the classification rules from these e-mails, A precise algorithm is then used.

I. INTRODUCTION

E-mails transfer any form of information between user systems having a proper internet connectivity. Unwanted emails in bulk, especially the commercial emails affects the storage of the mailbox memory. It would be difficult for user to delete each unwanted or unused mails manually. To handle this problem, with the increase in the problem of spam e-mails over the years numerous spam detection approaches has been developed. In general, all the e-mail messages are classified as "Ham" and "Spam". Ham messages are the intended or safe legitimate messages in a mail box; whereas Spam messages are the junk, unsolicited bulk or commercial messages in the mail box. This filtering or classification of email messages into Ham and Spam helps in separating them, so as to delete the spam messages through automation.

Usually there are several parameters or components which helps in identifying spam e-mails. An e-mail could be considered as Spam e-mail when it is associated with Bad grammar, Distorted images, Distorted symbols or logos, Bad links, Tempting offers and time based subscriptions that forces the users to subscribe immediately. Phishing is also considered as one of the dangerous cyber-crime which targets the individuals and tricks them to click on links or subscribe in order to steal the individual's data like login credentials of social accounts like Twitter, Facebook or internet banking details in the worst case scenario. Phishing e-mails are also considered as spam messages. This can also be manually prevented through unsubscribing e-mails, using safe e-mail readers/softwares like g-mail, yahoo, outlook etc., installing security softwares and keeping them updated all the time. But, it is not very easy to do as, sometimes important or useful information might be deleted and would not be possible to recover. Spam e-mails also include Spamvertised sites - e-mails that advertise products containing URLs that direct to other webpages, 419 Scams –spam e-mails wherea small initial payment in huge sum of money is offered to the users, Image spams – content present in a e-mail is displayed in the form of images.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 9 Sep. 2022, pp: 1436-1443 www.ijaem.net ISSN: 2395-5252

E-mail spam filtering is one of the frequently used process that helps in organizing all the e-mails based on a specified criteria. This process comes under automation as it automatically organizes all the e-mails based on prerequisites once they reach the mailbox server. These techniques of approach to spam filter does not follow any set of rules and regulations. So as to improve it further, it can be trained which helps in learning from previously grouped or classified spam or ham messages. This improvement is termed as Classification which includes the processes of Training and Filtering for a given dataset of e-mails.

There are also some problems that are associated with classification like Noise, Overfitting, Missing Values, Different forms of data. Noise is defined as the interference that occurs with reliability with which features are measured. Shadows, poor lighting conditions, images with blur, typing mistakes or intended misspellings to hide the spam messages from filters are considered as Noise. Overfitting occurs when there are too many attributes and relatively less observations, which identifies trained values perfectly but faces problem when classifying simple patterns of data and hence resolving makes the classifier comparatively more complex. Missing values are those in which dataset do not have information about all the features resulting in zero probability(Naïve Bayes Classifier) making it difficult to differentiate between the classes. Data may not always be in the same form. It may sometimes be the combination of images, text, videos etc. that cannot be used directly for the classifier. All these problems that are associated with classification should be taken into consideration to define a classifier perfectly.

Consumption space of recollection on servers which acquire added cost either to the user, provider or to the company although being of no usage altogether by the inception of Spam, considering a period of time and necessitating them to acquisition of additional storage. Furthermore, The extent of this storage compounds exponentially as millions of operators consumes the same e-mail client. It is very easy for the user to overlook or fortuitously delete emails which might be appropriate if regular emails are hustled along with spam. The reality of spam distresses an enterprise on all stages as critical communication on each level of an organization is reliant on e-mail.

II. LITERATURE REVIEW

The World Research Community displays huge curiosity on e-mail spam filtering which gained a rapid upsurge these past days. In this section, discussion of Similar reviews that are presented within the literature is done. Articulation of problems that are not yet addressed are surveyed in order to spotlight the conflicts within the review.

Usage of e-mails on both the professional and private stages and that they could also be well-thought-out as official documentsamongst individualsfor communication. Emailanalysis and data processing are going to be directed for several purposes like subject classification, spam detection and classification, etc. Revelation made clear that to filter the input file set by unsupervised filtering is utilized to overlook the utmost of prevailing researches. Maximum of prevailing practices that utilize additional features are limited to some substantial features of e-mails and might deliver significant result at most.

In 1998, by employing a Bayesian approach, Sahami et al. proposed a spam classification method focusing on unwanted e-mailsfiltering. The statistical classifier that works on independence of probability computation is a Bayesian classifier. With features of domainshown that accuracy are often improvedwhen Content of e-mail are considered. Later, many duplicated experimental results were found through his approach [1].

In 2005, to spot Anti-Spam Email, Zhan Chuan, LV Xian-liang has displayed an approachusing a new and improved e-mail filter based on Bayesian theory [4]. For representing word frequency, vector weights are used and attribute selection is adopted for counting on word entropy and a formula is deduced correspondingly. It was also demonstrated thattotal performances are improved deceptively by the filter.

In 2010, the need for effective spam filters have surged and also proved. V Christina et ai deliberated spam e-mails, spam filtering methods and their co-related problems [3]. Exploration of two key semantic methods by Man Qi et ai i.e. Sup-port Vector Machine (SVM) and Bayesian algorithms. In this paper Topical spam filters are conversed for determining spam messages that utilize semantic analysis of information [7].

In 2012,by applying rough set classifier, Perez-Diaz et al. evaluated on spam detection by means of diverse rule schemes of executionfor working out the simplest matching [6].

In 2013, for depth analysis of the receivede-mailsLalitha et al., listed the features for acknowledging the intended phishing mails along side the logosalso aspointers that are found that are unseen or hidden within the e-mails [2].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 9 Sep. 2022, pp: 1436-1443 www.ijaem.net ISSN: 2395-5252

In 2016, based notion for the classification and identification of spam emails, Tuteja and Bogiri have proposed Artificial Neural Network from manually formed dataset [8]. During this process, for feature extraction – K-means clustering is deployed; for the training of dataset - Back Propagation Neural Network is deployed;for the classification & identification of email spam - Feed Forward Neural Network is deployed. When compared to the results of pre-processing without using K-means clustering, that of with using K-means clustering are improved.

SVM-NB(Support Vector Machine combined with Naïve Bayes classifier) algorithm is proposed by Feng et al. for filtration in the e-mail spam messages [5]. For handling huge datasets, SVM algorithm and NB approach are combined and SVM is in a position to get a hyper-plane based separation between different features. They also compared SVM and NB reporting the better and efficient performance ofSVM-NB approach that is proposed.

III. PROPOSED METHODOLOGY

The proposed work reviews the models of Supervised Machine Learning which classifies the e-mails from the e-mail corpus as spam and ham in order to identify spam messages and mark them.
This methodology is presented as in the following Figure-1,

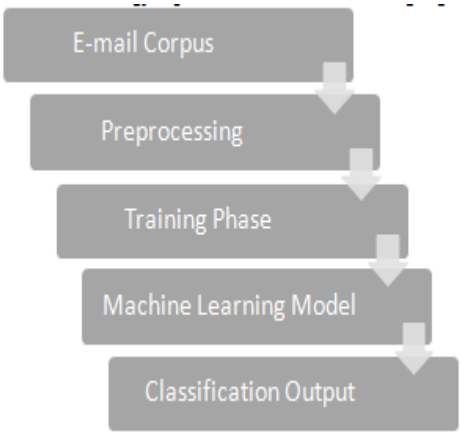


Figure-1

E-mail Corpus : All the messages in an e-mail are stored in order to process them. Basically it is a dataset or a database.
Preprocessing : The E-mail message that needs to be classified is initially pre-processed which include removal of null values, missing values and duplicate values.
Training Phase : The Data after preprocessing is split into two parts, Training and Testing. In Training Phase, the algorithm modifies the parameters for the model.
Machine Learning Model : The parameters are passed to the model, and based on the algorithm and process of model, it evaluates the given parameters and output is generated.
Classification Output : The output obtained from the classification model is then further classified as spam and non-spam.
A new testing phase can be added to the model to check the precision of the model. In this stage, based on predicted output and testing data, accuracy score is generated to define the perfection of model and compare with other models.

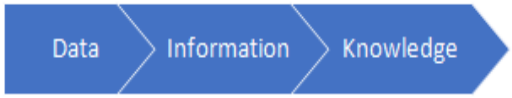
IV. CONCEPTS USED

Machine Learning is a field of scientifical study of algorithms and mathematical or statistical models that a computer utilize to attain capability of learning or to perform a specific task without using explicit set of instructions, relying on patterns and inference instead. It is a subset of a broad field of Artificial Intelligence which allows machines and computers, act or perform certain activities like human does. Machine Learning comes into application in many scenarios like Spam Detection, Speech and Image Recognition Systems, Medical Diagnosis, Prediction Systems etc.. It helps in reducing human effort, hence making the tasks easy to be performed with the help of a machine.There are a lot of algorithms that could be used in e-mail filtering, which are broadly studied by the approach of Machine Learning.This includes K-nearest neighbour(KNN),Naïve Bayes Theory, Support Vector Machines(SVM), Artificial Neural Networks(ANN), Rough Sets Classifiers, Artificial Immune System and so on.

Broad division of Machine Learning is made into three major categories, depending on the nature of learning. They are Supervised Machine Learning, Unsupervised Machine Learning, Reinforcement Learning. Supervised Learning provides the system with certain inputs and corresponding outputs where a general rule is generated that maps input to its corresponding output(example: Spam detection, fraud detection, image recognition). Unsupervised Learning is where outputs are not defined, allowing the system to find a pattern from the given input(example: grouping fruits based on size, shape or color).Whereas in Reinforcement Learning, A computer program interacts with an environment to reach certain goal and it do not have any prior knowledge about the target(example: robotic systems, learning to drive a vehicle).

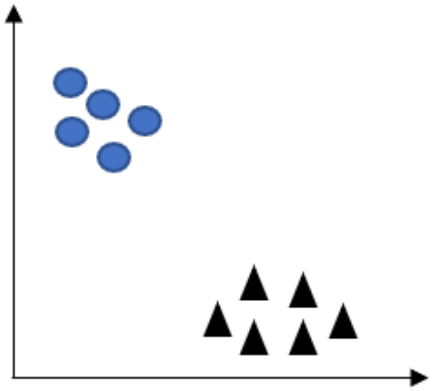
International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 9 Sep. 2022, pp: 1436-1443 www.ijaem.net ISSN: 2395-5252

Machine Learning includes a lot of pre-processing required for an algorithm to work more efficiently. Initially, Data(any unprocessed text, value, fact, sound or a picture) is converted to Information(interpreted and manipulated data) and further made useful by providing it in the form of Knowledge(further inferred resulting in concept building).

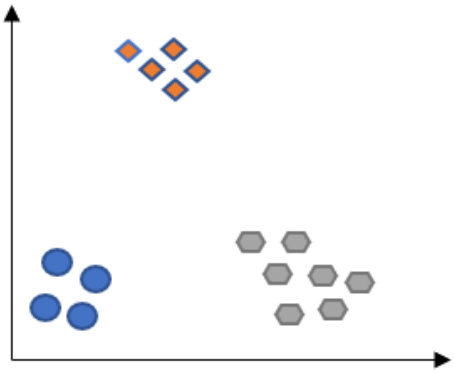


Data is split to perform several actions like Training, Testing and Validation. Processing of Data is done through the steps of Collecting, Preparing, Input, Processing, Output, Storage.Data Processing, Data Cleaning takes place includesHandling missing data,Exclusion ofunessential observations, Fixing Structural errors, Managing Unwanted outliers.

As Supervised Machine Learning models are used by us for e-mail spam detection, Classification is majorly used for spam detection, as the name implies, grouping or classifying a similar object into categories based on the training dataset obtained. Classification can further be divided into two sub-categories i.e. Binary Classification – Categorizing data into two distinct classes, Multiple Classification – Categorizing data into multiple(more than 2) subclasses.



a. Binary Classification



b. Multiple Classification

Some of Supervised Machine Learning Techniques that are frequently used for e-mail spam detection are:

- K-Nearest Neighbour(KNN)
- Naïve Bayes Theory
- Artificial Neural Networks (ANN)
- Support Vector Machine (SVM)
- Rough Set Classifier

1. K-Nearest Neighbour (KNN) :

The K-nearest neighbour (KNN) classifier in which the usage of training documentsfor comparison as an alternative of a particular categoryrepresentation hence called asan instance-based classifiertaken into account, like the categoryprofiles employed by other classifiers. In KNN, there's no realtraining phase. The k most similar documentsor neighbours are found when a replacement documentmust be categorized and of them an outsized enough proportion are assigned to a particular category, the new document is additionally allocated to the present category, else not. Additionally,using traditional methods of indexingfinding the nearby neighbours are often fastened. We glance at the category of the messages that are closest thereto, to decidewhether a message is spam or ham. The comparison between the vectors may be a real-time process is the often thought of the k-nearestneighbour algorithm:

Assumption : In a given dataset, instances with similar properties exists close to each other.

Training : Split the training dataset and store it.

Filtering : For a given message, determine k nearest neighbours for each attribute within the training dataset. Classify the spam messages among neighbours as spam, else classify them as ham.

K-Nearest Neighbour being an example-based classifier, consumes less computational time in training and more computational time in testing.

Algorithm :

Step-1 : Load the Training Data

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 9 Sep. 2022, pp: 1436-1443 www.ijaem.net ISSN: 2395-5252

Step-2 : For each individual test instance, evaluate the Distance Metric (distance from each training instances used) by calculating the Euclidean Distance,

$$D(x,y) = \sqrt{(\sum_{i=1}^{n} |xi - yi|^2)}$$

Step-3 : Find the k-neighbours with the nearest (minimum) distance

Step-4 : Consider the label which have major votes among the given dataset labels to decide the label of test instance.

PROS :

- Output obtained is of high accuracy for small datasets
- Takes all the features present in the dataset into consideration

CONS :

- Computes all the training instances per test instance during classification, resulting in high time complexity during the testing phase, further increasing the computational cost
- Require large amount of memory

2. Naïve Bayes Theory :

Naïve Bayes algorithm may be a statistical machine learning based Bayes approach that usually have the strong independenceproperties, probability distribution and skill to tacklehuge datasets. In NB, from the distribution of datasetprobability distribution is evaluated. Bayes decision rule is employed to assign a class in classification problem. Classes having the highest value of posterior probability are chosen by the classifier as defined by the bayes decision rule. The posterior probabilities are often evaluated with the following equation,

Based on Bayes Theorm for Conditional Probability, Probability that a given set of features (x1,x2,…..,xn) are enclosed in a vector V belonging to a category or a class C is given by,

$$P\left(\frac{C}{V}\right) = \frac{P(C).P\left(\frac{V}{C}\right)}{P(V)}$$

$$P\left(\frac{S\_}{V}\right) = \frac{P(S\_).P\left(\frac{V}{S\_}\right)}{P(S\_).P\left(\frac{V}{S\_}\right) + P(H).P\left(\frac{V}{H}\right)}$$

Assumption : The values of a specific feature is independent of all the other features given in that class.

Training : Parse each e-mail message into its respective tokens, then a probability is generated for each token and values of spam probability are stored.

Filtering : For an each e-mail message, categorize them into spam and ham considering a threshold value to define spam content.

GAUSSIAN NAÏVE BAYES FILTERING:

Assumption : Consider that continuous values follow Gaussian Distribution.

Training : For training dataset, Segment the data by class and compute the mean(μ) and variance(σ^2) of all the values present in for each class.

Filtering : For test instances with attribute value(v), probability that the instance belongs to class 'C' is

$$P\left(\frac{X = V}{C}\right) = \left(\frac{1}{\sqrt{2\pi\sigma 2}}\right) e^{\left(\frac{(v-\mu)2}{2\sigma 2}\right)}$$

PROS :

- Fast Training – Computes mean and variance of training data
- Simple Statistical approach – easy to implement

CONS :

- Might not be able to hold well when data is correlated or assumption of data independence fails
- Affected by zero probabilities (occur when product of individual probabilities = 0; due to missing values)

3. Support Vector Machine (SVM) :

Support Vector Machine (SVM) is grounded on the principle of Structure Risk Minimization which aims at finding the hyperplane that divides the two given classes perfectly. Points lying on the hyperplane are called Support Vectors and these are the only points used in the decision function.

The concept of decision planes that outline decision boundaries supports Support Vector Machines. A group of objects having non-identical class memberships is separated by a choice plane, and an optimal hyperplane with the maximal margin used in separating two classes is found by the SVM modelling algorithm, which involves simplifying the subsequent optimization problem.

A Cross Validation is a typical process that is conducted on the training dataset. Cross validation additionally necessitated to estimate the generalization capability on new samples that aren't within the training dataset. A cross validation randomly splits the training dataset into k precisely equal-sized subsets called k-fold, in which one subset is left out, and a classifier is built on thesamples remaining, then classification performance is evaluated on the unused subset. This procedure is recurred k times for each and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 9 Sep. 2022, pp: 1436-1443 www.ijaem.net ISSN: 2395-5252

every subset to obtain the cross validation performance over the whole training dataset. A little subset are often used to minimize computing costsfor cross validation, If the training dataset is large. the subsequent algorithm are often utilized in the classification process.

Training : From all the samples of training set that require classification, find k nearest neighbours for them. Obtain the decision points and train the SVM model.

Filtering : Classify all the attribute points from the obtained model on either side of hyperplane and outputs the result.

PROS :

- SVM, for high dimensional spaces is very influential.
- Memory efficient – SVMin its decision function utilize the subset of training points.

CONS :

- SVM may not be effective when the number of features are comparatively greater than the number of samples.
- SVM does not output direct probability values, hence need cross validation.

4. Artificial Neural Network :

An artificial neural network (ANN), also called in general a "Neural Network" (NN), which may be a computational that model supports biological neural networks. Organized assortment of artificial neurons are comprised within. A Neural Network that made by a human is an adaptive structurewhichmodifies its model supported information that streamsalong the synthetic network all throughthe learning phase. The ANN is predicated on the principle of knowledge by example. There are, however the 2 classical quite the neural networks, perceptron and therefore the multilayer perceptron. Here we'll specialise in the perceptron algorithm. the thought of the perceptron is to seek out a feature vector in linear function i.e.

$$f(x) = W.T(x) + b\ ;$$

$$f(x) > 0\ ;\ \text{class1}, \qquad f(x) < 0\ ;\ \text{class2}$$

Where $W = (W1, W2, \ldots . Wm)$ is that the vector of coefficients (weights) of the function, and b is the so-called bias. If we denote the classes 1 and 2 by numbers +1 and -1, we can state that we are looking for a choice function, $d(x) = \text{Sign}(W.T(x) + b)$. With an iterative algorithm, the perceptron learning is completed. It starts with the arbitrarily chosen parameters $(w0, b0)$as the choice and are updates iteratively. A training sample (x, c) is chosen such that the present decision function doesn't classify it correctly (i.e.$\text{Sign}(Wn.x + bn \neq c)$ on the nth iteration of the algorithm. The parameters 'wn' and 'bn' are then updated using the rule.

$$W(n+1) = W(n) + cx$$

$$b(n+1) = b(n) + c$$

This Algorithm terminates when classification is done correctly on all the training samples by the decision function that is found.

Training : Values of the training dataset are initialized and their weights are updated supported on the specified required output. This is often repeated until the requirement is satisfied, hence completing the training.

Filtering : For a given e-mail message, determine its class.

PROS :

• Neural networks are flexible and should be used for both regression and classification problems. Any data which can be made numeric are often utilized within the model, as neural network could also be a mathematical model with approximation functions.

• Neural networks are good to model with nonlinear data with sizable amount of inputs; as an example, for instance , images. it's reliable in an approach of tasks involving many features. It works by splitting the matter of classification into a layered network of simpler elements.

• Neural networks are often trained with any number of inputs and layers, also predictions are pretty fast

• Neural networks toil finest with extra data points

CONS :

• Neural networks are black boxes, meaning we cannot skills much each experimental independent variable is influencing the dependent variables.

• It is computationally very expensive and time consuming to teach with traditional CPUs.

• Neural networks depend tons on training data. This leadsto the matter of over-fitting and generalization. The mode relies more on the training data and should be tuned to the info .

5. Rough Set Classifier :

Computing the reductions of data systems is a great ability featured by Rough Set Classifiers. Attributes irrelevant to the target concept (i.e. decision attributes), and a few redundant attributes might be present in the data system. To attain simple useful knowledge from it,Reduction is required,Which is a minimal subset of condition attributes that are referenced to the decision attributes. The Rough set scheme is as follows.

i. The first thing that we would try with the incoming mails is picking foremost appropriate attributes to be further used for classification. Later, the dataset input is

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 9 Sep. 2022, pp: 1436-1443 www.ijaem.net ISSN: 2395-5252

transformed into a choice system, which further splits into training and testing datasets. The training dataset induces a classifier within which is applied to testing dataset to evaluate performance estimation. For Training the dataset, step-2 and step-3 are followed.

ii. Boolean reasoning need to finish the discretization strategies as the decision system has the real value attributes.

iii. For obtaining the decision rules, genetic algorithms should be utilized. Proceed with step-4 for the testing dataset.

iv. For employing equivalent cuts that are computed from step-2, discretize the testing dataset.Make sure, each and every new object in the testing dataset need to match with the principles generated in step-3.

V. RESULTS AND DISCUSSION

Machine Learning algorithms play a crucial role when it comes to spam classification. There are several new algorithms that are developed from the old one's or their combination. Five major machine learning models that are used in spam classification are discussed in this paper.

E-mail messages consists of numerous parts: header, body etc. Header contains the fields in the mail like 'From', 'Subject'. Subject consists most of the information which is generally used to classify as spam or ham; whereas From is used to know about the sender and to mark the sender as spam if required, so that all the e-mail messages from that sender can be directed to the spam folder without any further classification process required. Body is main part of the e-mail message which defines the structure of the message for proceeding with steps of preprocessing. Several features in the Body are selected to define or categorize the words as spam which further defines the message as spam message.

While consideration of methods is done, they are choosed based on the features selected or how the message should be classified. Every classification algorithm has its own advantages and disadvantages when parameters like computational time, computational cost, memory allocated, etc.
We consider three parameters to define the performance of an algorithm,

i. Accuracy : The e-mails that are properly classified and categorized per all e-mails considered based on the accuracy score. It defines how accurately the algorithm works.

ii. Spam Recall : The spam e-mails that are properly classified and categorized as spam per all spam e-mails considered is Spam Recall.

iii. Spam Precision : The Spam Precision defines the percentage of related spam e-mails identified among all the e-mails. Shows how many e-mails classified and categorized as spam are actually spam.

Table-1

ALGORITHM USED	ACCURACY (in %)	SPAM RECALL (in %)	SPAM PRECISION (in %)
K-Nearest Neighbour	96.20	97.14	87.00
Naïve Bayes	99.46	98.46	99.66
Support Vector Machine	96.90	95.00	93.12
Artificial Neural Network	96.83	96.92	97.75
Rough Set Classifier	97.42	92.26	98.70

As observed in the above Table-1, All the Performance metric parameters(Accuracy, Spam Recall, Spam Precision) are highest for Naïve Bayes.

VI. CONCLUSION

Through this study in the paper, we learned about detecting the spam messages in e-mails through different approaches of classification algorithms by machine learning. This review justifies the working and functionality of the algorithms along with their advantages and disadvantages based on numerous considered parameters. To solve the problem of spam e-mails through machine learning classifiers, several attempts has been made by many researchers. This also became a leverage in producing new loopholes for spam e-mail generation. Detection of spam e-mail messages have evolved from filtering to classification. Besides, there are numerous amount of algorithms from which some of the major algorithms are looked into. This paper is presents the review based on several challenges based in Spam filtering and classification when a particular algorithm in considered in specific. Major studies and researches that are developed based on several challenges have been discussed in the literature review. Some of the open research problems also includes usage of these algorithms that has been thoroughly reviewed and performance metrics for an algorithm is evaluated form accuracy, spam recall and spam precision,

In brief, this paper discusses how a spam detection and processes of filtering and classification works, current trends of spam, how the approach of machine learning field helps in spam detection process, how a general supervised machine learning classification algorithm works, how a specific algorithm classifies the e-mails into substituent spam and ham messages, the parameters in which a particular algorithm is efficient and in which it isn't. Through this review, selection of a particular algorithm can be made based on the features considered in detecting a spam e-mail, Also is helpful in developing a hybrid algorithms through combination of a algorithms as their peer review is made. As observed from all the models of classification in the field of machine learning, every method that is considered has its own pros and cons. So, for an efficient algorithm to be developed that performs at best even when any parameters like evaluation time, acquaintance cost, memory of allocation etc. Therefore, Hybrid Algorithms seems to be the best and feasible solution for Spam detection in e-mails.

REFERENCES

[1] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105)

[2] Lalitha, Mrs P., and SumalathaUdutha, "New Filtering Approaches for Phishing Email," International Journal of Computer Trends and Technology (IJCTT) 4.6 (2013): 1733- 1736

[3] V Christina., "A study on email spam filtering techniques", international Journal of Computer Applications, Vol. 12- No.1, 2010.

[4] Zhan Chuan, LU Xian-Iiang, ZHOU Xu, HOU Meng-shu, "An Improved Bayesian with Application to Anti-Spam Email ", Journal of Electronic Science and Technology of China, Mar. 2005, Vol.3 No.1

[5] Feng, Weimiao, Jianguo Sun, Liguo Zhang, Cuiling Cao, and Qing Yang. "A support vector machine based naive Bayes algorithm for spam filtering." In Performance Computing and Communications Conference (IPCCC), 2016 IEEE 35th International, pp. 1-8. IEEE, 2016.

[6] Pérez-Díaz, Noemí, et al., "Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification," Applied Soft Computing 12.11 (2012): 3671- 3682.

[7] Man Qi, Mousoli, R, "Semantic analysis for spam filtering", international Conference on Fuzzy Systems and Knowledge Discovery, VoI.6,Pg.2914-2917, 2010.

[8] Tuteja, Simranjit Kaur, and Nagaraju Bogiri. "Email Spam filtering using BPNN classification algorithm." In Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on, pp. 915-919. IEEE, 2016.