Received: November 14, 2017

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Machine Learning based Spam E-Mail Detection

Priti Sharma[1] **Uma Bhardwaj**[1]*

[1]*Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India*
* Corresponding author's Email: umabhardwaj90@gmail.com

Abstract: Spam email is one of the biggest issues in the world of internet. Spam emails not only influence the organisations financially but also exasperate the individual email user. This paper aims to propose a machine learning based hybrid bagging approach by implementing the two machine learning algorithms: Naïve Bayes and J48 (decision tree) for the spam email detection. In this process, dataset is divided into different sets and given as input to each algorithm. Total three experiments are performed and the results obtained are compared in terms of precision, recall, accuracy, f-measure, true negative rate, false positive rate and false negative rate. The two experiments are performed using individual Naïve Bayes & J48 algorithms. Third experiment is the proposed SMD system implemented using hybrid bagged approach. The overall accuracy of 87.5% achieved by the hybrid bagged approach based SMD system.

Keywords: Correlation based feature selection, Spam filtering, Hybrid bagged approach, J48 algorithm, Naïve bayes, Text mining.

1. Introduction

Email system is one of the most effective and commonly used sources of communication. The reason of the popularity of email system lies in its cost effective and faster communication nature. Unfortunately, email system is getting threatened by spam emails. Spam emails are the uninvited emails sent by some unwanted users also known as spammers [1] with the motive of making money. The email users spend most of their valuable time in sorting these spam mails [2]. Multiple copies of same message are sent many times which not only affect an organisation financially [3] but also irritates the receiving user. Spam emails are not only intruding the user's emails but they are also producing large amount of unwanted data and thus affecting the network's capacity and usage. In this paper, a Spam Mail Detection (SMD) system is proposed which will classify email data into spam and ham emails. The process of spam filtering focuses on three main levels: the email address, subject and content of the message [4].

All mails have a common structure i.e. subject of the email and the body of the email. A typical spam mail can be classified by filtering its content. The process of spam mail detection is based on the assumption that the content of the spam mail is different than the legitimate or ham mail. For example words related to the advertisement of any product, endorsement of services, dating related content etc. The process of spam email detection can be broadly categorized into two approaches: knowledge engineering and machine learning approach [5]. Knowledge engineering is a network based approach in which IP (internet protocol) address, network address along with some set of defined rules are considered for the email classification. The approach has shown promising results but it is very time consuming. The maintenance and task of updating rules is not convenient for all users. On the other hand, machine learning approach does not involve any set of rules and is efficient than knowledge engineering approach [6]. The classification algorithm classifies the email based on the content and other attributes. For most of the classification problems the process

DOI: 10.22266/ijies2018.0630.01

of feature extraction and selection is very important. Features play a vital role in the process of classification. In this paper, a correlation based feature selection (CFS) [7] method is used for feature extraction. The CFS approach extracts the best features from the pool of features for efficient classification results. In order to remove the drawbacks of current model a novel hybrid bagged technique is introduced in the proposed spam mail detection (SMD) system. Fig. 1 presents the basic process of email filtering.

The proposed spam mail detection system is inspired from the effectiveness of machine learning approach. In spam mail detection system, initially email data is collected. The email data collected is raw and unstructured in nature. In order to reduce the computations and to obtain accurate results, email data needs to be pre-processed. The data is pre-processed by removing stop words, stemming and word tokenization is also performed to acquire valuable information. Then, CFS based i.e. correlation based feature selection is performed to get the best selected features from the pool of features. The pre-processing step reduces the dimensionality of data and features in the form of bag of words are then extracted. For the classification a bagged hybrid approach (which is combination of Naïve Bayes classifier and J48) is used in order to make the classification stronger and more accurate. The dataset is randomly divided into different sets and serves as input to each classification algorithm. The bagging approach combines the classification results of the two machine learning algorithms to evaluate the final classification result.

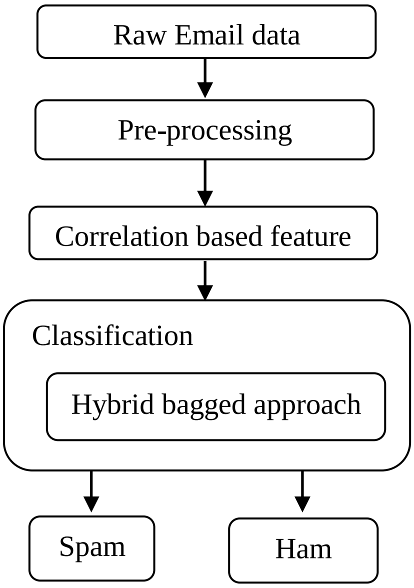

Figure. 1 Basic process for email filtering

The rest of the paper is structured in the following manner: Section II shows the work of authors related to email filtering. Section III includes the preliminaries of proposed model. In section IV, the modules of the proposed Spam Mail Detection are explained along with working example. Section V illustrates the calculated experimentation results and Section VI concludes the paper.

2. Related work

Email system is one of the most common and popular communication systems. Organisations from all over the world are making their efforts in order to identify the spam mails. The work of authors to identify the ham and spam emails is discussed here. Table 1 illustrates the comparative work of authors by stating the classification techniques, dataset, feature extraction approaches and drawbacks.

In order to classify the email as ham or spam, a filtering technique is required for its classification. Mohamad and Selamat [8] have proposed a spam email filtering system using two different features selection methods to classify the emails. They have considered English and Malay email dataset and after the pre-processing of the dataset features are selected using TF-IDF and rough set theory method. Then machine learning technique is applied for the classification purpose showing some reasonably good results. Another machine learning algorithm based work for the classification of email data was presented by Harisinghaney et al. [9]. The algorithmic implementation includes KNN, Naïve Bayes and DBSCAN algorithms and showing effective results when the algorithms are applied on pre-processed data.

Further, Youn and Mcleod [10] proposed an ontology based email filtering method. The considered dataset is classified using J48 decision tree based algorithm. A RDF language based ontology is created by Jena in order to test the results obtain after the classification.

Authors have also adapted the optimization techniques. Faris et al. [11] have used feed forward neural network based method to detect the spam emails and to optimize the results as well. The neural network is trained with the help of Krill Herd algorithm. The pre-processed dataset is equally divided into two halves for the training and testing purpose. The optimized classification results obtained from neural network are compared with other optimization algorithms like Genetic algorithm and Back propagation. The experimental results shown by Kill Herd algorithm are more accurate than the other two algorithms. Another optimization based system is proposed by Al-Shboul et al. [12] for the detection of spam mails. The authors have considered a hybrid approach for the email filtration process. In the first phase, Particle Swarm Optimization based algorithm is considered in order to select the best and optimized features. In the second phase, Random forest algorithm is trained with the selected features form the previous phase in order to classify the email dataset into ham and spam emails.

There are various techniques for the classification of emails into spam and non-spam emails. Tuteja [13] has presented the review on different techniques for the spam and non-spam email classification. The author has presented different classification algorithms like Support Vector Machine (SVM), Naïve Bayes (NB) classifier, Neural Network (NN), J48 Decision Tree based classifier for the classification of the emails. In this paper, the author Tuteja [14] has proposed a system based on neural network for the classification of the emails into spam and non-spam emails. A review on current trends and techniques for email classification has also been presented by Mujtaba et al. [14]. The author has suggested three levels in every classification technique. The pre-processing state which includes conversion of words into tokens, removal of stop words etc. The second phase is learning. In this level, the feature set is prepared which is required for the classification. Here features are extracted and then selected. The third and final level is classification level. In this level, the classifier is considered which classifies the email into its respective class. Different algorithms are considered for the classification like, support vector machine, logistic regression, regression trees, and random forest etc to name as few.

Recently, machine learning algorithms are integrated with some other algorithms in order to acquire more effective results. Ajaz et al. [15] have proposed a spam email detection system by considering a hybrid approach of Secure Hash algorithm and Naïve Bayes machine learning classifier. The feature set is considered with the help of Naïve Bayes classifier. With the help of SHA (Secure Hash algorithm) a function is generated which considers the email in form of message M. The message M is further classified into two categories S and L. where S stands for the spam message and L indicates genuine message. The work proposed by Ajaz et al. [15] has shown effective results.

3. Preliminaries

In this section, the preliminaries of Naïve Bayes classifier and J48 decision tree algorithm are discussed.

3.1 Naïve bayes classifier

Naïve Bayes classifier is based on Bayes theorem with an assumption of strong independence [16]. The classifier is a probability based classifier which computes the class probabilities of the given instances. The probability set is calculated by computing the combinational and frequency values of the data set. The class probability which is nearest to the rear end will be picked by the classifier. The Naïve Bayes classifier is a multiclass classifier and works efficiently with supervised learning approach. The concept of the Naïve Bayes classifier is explained with the help of Eq. (1).

$$P(y|x) = \frac{P(x|y)\ P(y)}{P(x)} \quad (1)$$

Here, x is the set of feature vectors (x_1, x_2, $x_3, \ldots .x_n$) and y stands for the class variable with $m$possible outcomes (y_1, y_2, $y_3, \ldots .y_n$). $P(y|x)$ is the posterior probability which depends on the likelihood of the feature set or attribute value belonging to particular class $P(x|y)$, $P(y)$ is the prior probability and $P(x)$ is the evidence depending on the known feature variables.

3.1.1. Multinomial naïve bayes

The Naïve Bayes classifier used in the proposed work is Multinomial Naïve Bayes classifier. In this classifier, data is represented in the form of counts of word vectors [17]. The parameterised distribution by vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each y class, where n represents features and θ_{y1} is the probability $P(x_i|y)$ of feature i in particular sample belonging to class y. Eq. (2) shows Multinomial Naïve Bayes classifier mathematically.

$$\hat{\theta} yi = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad (2)$$

Let us consider an example to show the working of Naïve Bayes classification algorithm. Assume that the email contains the word “profit”. People who operate on email system for their communication would know that this email is likely to be a spam email. The spam mail detection system implemented using Naïve Bayes classification algorithm computes the probability to classify the email with word “profit” using the Eq. (3).

Table 1. Author's work at a glance

Author	Email Dataset	Feature Extraction Approach	Classification Technique	Drawbacks
Mohamad and Selamat [8]	• English and Malay email dataset	• TF-IDF • Rough set theory	• Machine learning	The characteristics of image based emails like shape, color, texture etc. are avoided and only emails with embedded text are considered.
Harisinghaney et al. [9]	• Enron Corpus	• Black and White listing • Google open source library Tesseract	• K-Nearest Neighbour • Naïve Bayes • DBSCAN algorithms	The proposed approach is time consuming and recognizes text characters limited to only certain number of fonts.
Youn and Mcleod [10]	• Dataset from UCI machine learning lab	• Ontology based	• J48 decision tree	The proposed spam email filtering system only works for the emails in a comma separated value (CSV) format.
Faris et al. [11]	• SpamAssassin	• Krill Herd algorithm	• Feed forward neural network	The algorithm feed forward neural network is trained after each iteration.
Al-Shboul et al. [12]	• SpamAssassin	• Particle Swarm Optimization	• Random forest algorithm	The email dataset considered for the experimentation is imbalanced data consisting of approximately 25.6% of the spam emails.
Tuteja [13]	• Spambase	• 11 binary features extracted from Header and body of email	• Artificial neural network	The study of the classification algorithms did not state the advantages and disadvantages of particular algorithm for the spam email filtering system.
Mujtaba et al. [14].	• Phishing Corpus	• Bag of words	• Machine learning	The study presented the different learning approaches for email classification but did not state the different tools, feature selection and reduction methods for email classification.
Ajaz et al. [15]	• Online available spam emails	• Naïve Bayes	• Secure Hash algorithm	The proposed secure hash method filter the email data but unable to dissipate the misuse of storage resources and network bandwidth.

$$p(s|w) = \frac{p(W|S)\, p(s)}{p(W|S)\, p(s) + p(W|h) p(h)} \tag{3}$$

Here w representing the word "profit" h and s represents email classes ham and spam respectively. The probability of email belonging to spam class containing word "profit" is *p(s|w)* depends on the overall probability of any email belonging to spam class *p(s)*, the probability of occurrence of word "profit" in spam emails *p(w|s)*,the overall probability of any email belonging to ham class *p(h)* and the probability of occurrence of word "profit" in ham emails *p(w|h)*. Algorithm 1 presents the training and applicability of the Multinomial Naïve Bayes classification algorithm. In this algorithm, C is set of all classes, c is the computed class, N is number of documents and N_c is number of documents in the class c.

3.2 J48 classifier

The J48 classifier is a decision tree classifier based on the concept of entropy. It is a multiclass classifier forming decision trees of the training data. The decision tree generated using J48 depends on the training data attribute values for the classification of the new data item. J48 follows the concept that by splitting the data into multiple sets, each feature attribute of data can be used to form a

Algorithm 1

```
Train_multi_NB(C,D)
{
count total number of documents N
extract vocabulary voc from the document D
for each c ∈ C do
count number of documents in class c , $N_c$
prior [c] = $N_c$ / N
concatenate text $text_c$ of all documents in class c,
    for each t ∈ V do
        count tokens of term t from $text_c$ , $T_{ct}$
        conprob [t][c] = $\frac{Tct+1}{\sum_{t'}(Tct'+1)}$
return (V,prior,conprob)
}

Apply_multi_NB(V, C, prior, conprob, d)
{
Extract tokens W from document d and vocabulary V
for each c ∈ C do
score[c] = log prior[c]
    for each t ∈ W do
        score[c] += log conprob [t][c]
return $max_{c \in C}$ score[c]
```

decision [18]. The algorithm works recursively until each data attribute is processed and categorized i.e. the features extracted with the help of this algorithm are the best possible features belonging to the particular class data. The algorithm has few considerations which are listed as follows:

- When the instances of the already considered unseen classes are encountered, then a decision node higher to the tree is created by the algorithm.
- In case when the sample data belongs to single class, the algorithm creates decision tree with leaf node and ask to consider that particular class.
- A decision node is created above the existing tree using expected values if the attributes or feature values do not provide any gain in information**.**

A decision tree consists of root node, internal nodes and leaf nodes. Internal nodes represent the conditions applied on attributes or features whereas leaf nodes represent the class. Fig. 2 presents example of a typical decision tree. Further Algorithm 2 is presented for J48 decision tree algorithm.

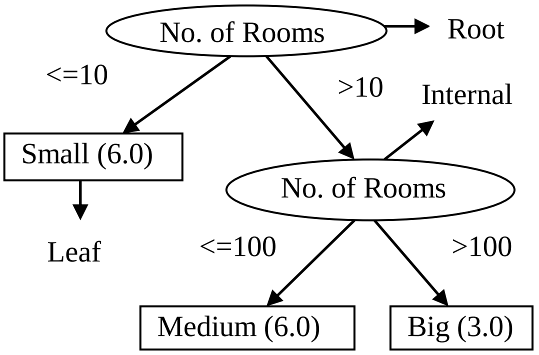

Figure. 2 Example of decision tree

Algorithm 2

```
build_decision_tree(*D)
{
create root node T and label it with splitting attribute
add arc to root node T for each split label
for each arc do
 D= database created by splitting feature attributes to D
  if stop criteria met then
     Create leaf node T'
     Label leaf node with appropriate class
  else
     T' = build_decision_tree(D)
  T = add T' to arc
}
```

In Algorithm 2, *D* is the set of training data and *T* is the decision tree. In this algorithm, splitting criteria is a feature selection method that splits the dataset items into particular individual classes.

4. Spam mail detection system

This section represents the workflow of the Spam Mail Detection (SMD) System for the classification of emails into ham and spam emails. The SMD system consists of strong classification abilities introduced with the concept hybrid bagged approach. The feature selection method is performed with correlation based feature selection and performed classification with novel hybrid bagging approach. The bagging approach is a hybrid approach where decision tree based J48 algorithm and Naïve Bayes Multinomial classifier is the classification purpose. The flow chart of the SMD system for email classification is presented in fig. 3.

The SMD system classifies the email into spam and ham emails. The text based email dataset considered is initially pre-processed for efficient feature extraction. The classification approach considered is a hybrid bagged approach. The SMD system consists of four modules (sub-sections 4.1 to 4.4) of Email Dataset Preparation, pre-processing of data, feature selection and hybrid bagged approach. Also a working model is explained in sub-section 4.5.

4.1 Email dataset

An email dataset is prepared for the Spam mail detection system. Different emails are randomly collected from Ling spam dataset [19]. The dataset consists of total number of 1000 emails consisting of both ham and spam emails for the classification purpose. Since the approach considered is bagging approach, the dataset is divided into sets for each classifier algorithm. Two sets of 500 emails each have been prepared. Out of which 300 emails each are used for training purpose for both the Naïve Bayes and J48 algorithm and 200 each for the testing purpose. The statistics of dataset is shown in table 2.

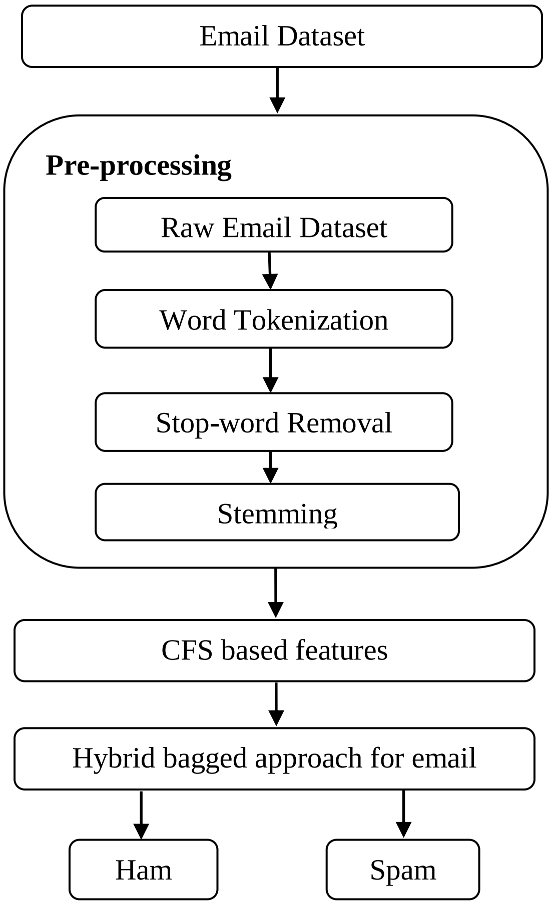

Figure. 3 Spam mail detection system for email classification

Table 2. Email dataset statistics

Email Dataset	**Naïve Bayes**	**J48**	
No. of Ham Emails for training	150	150	
No. of Spam Emails for training	150	150	
No. of Ham Emails for testing	100	100	
No. of Spam Emails for testing	100	100	
Total Emails	500	500	**1000 Emails**

4.2 Pre-processing of dataset

The email dataset considered is raw in nature. So it needs to be pre-processed before further consideration. The pre-processing phase consists of three steps. Initially the tokenization of the text data is done. The sentence is split into words known as tokens. From the tokenized words, stop words are removed. Stop words are unwanted words having no linguistic meaning. A text file of approximately 670 stop words is manually prepared and words are removed from the text at the time pre-processing. The third step in the pre-processing module is the stemming. The process of stemming reduces the word to its base word. Stop word removal and stemming are important steps in the pre-processing phase as they help to reduce the search space for efficient feature extraction and selection.

4.3 Feature selection

Features play an important role in any of the classification system. SMD system works with assumption that spam mail differs than the ham mail in terms of its content. The feature set contains different features like alphanumeric words, language, grammatical or spelling errors, inappropriate words (words related to advertisement of products/services, dating, adult words etc), frequency count, document length etc. In SMD system correlation feature selection (CFS) method is used. CFS only identifies the best features among the pool of features which are helpful in improving the performance of the system. Correlation based feature selection method works on the assumption that, "Good feature subsets contains features highly correlated with the classification, yet uncorrelated to each other" [7].

Initially text data with feature set is considered as bag of words. The term frequency method is considered to show the number of words per document. The frequency of all words is calculated and words with frequency below a threshold value are eliminated. This method indicates the usefulness of the words and also reduces the search space. The obtained feature set is further reduced using correlation based feature selection method.

Correlation based feature selection method only selects the feature set which are most related to the particular class. If f is the feature set with k number of features and c is number of classes then Eq. (4) presents the mathematical formulation of correlation based feature selection method.

$$CFS = max_{S_k}\left[\frac{r_{cf_1},r_{cf_2},r_{cf_3},\ldots,r_{cf_k}}{\sqrt{k+2\left(r_{f_1f_2}+\ldots r_{f_if_j}+\ldots r_{f_kf_1}\right)}}\right] \qquad (4)$$

Here r_{cf} is the average of feature-class correlation, r_{ff} is the average of feature-feature correlation.

4.4 Hybrid bagged approach

The fourth and last module is the classification module. A hybrid bagged approach with a combination of the decision tree based J48 algorithm and Naïve Bayes Multinomial classifier is considered for classification. Bagging approach also called as bootstrap aggregating approach decreases the variance by considering the combinations of the multiple repeated sets of the same dataset. In this approach, multiple models are generated by dividing the email dataset is randomly divided into separate sample email datasets: SED1 and SED2. Each sample email dataset is considered to train individual classifier. The overall system's result is the average of the result of the two classification algorithms. J48 algorithm and Naïve Bayes are used for the multi class learning and for the classification. The classification result considered is the average of the predicted values. Fig. 4 presents the concept of bagging.

4.5 Working of SMD

The detailed description of the modules of Spam Mail Detection (SMD) system is given with the help of the following example. An example of a random email is considered to explain the step-wise working of Spam Mail Detection system. The SMD system takes input in the form of an email and delivers the output in the form of either spam or ham as shown in table 3.

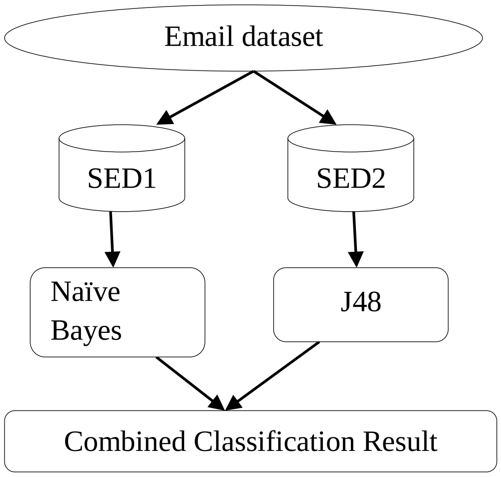

Figure. 4 The concept of bagging

Table 3. Working of SMD

Input	Subject: A guaranteed return of 70% on investment . . . $300 million increase in the economic growth in a year. After breakup of the largest monopoly of America earn some profit. Over 70% of returns annually. For complete details just click here
Tokenization	'Subject'':' 'A' 'guaranteed' 'return' 'of' '70%' 'on' 'investment' '. . .' '$300' 'million' 'increase' 'in' 'the' 'economic' 'growth' 'in' 'a' 'year' '.' 'After' 'breakup' 'of' 'the' 'largest' 'monopoly' 'of' 'America' 'earn' 'some' 'profit' '.' 'Over' '70%' 'of' 'returns' 'annually' '.' 'For' 'complete' 'details' 'just' 'click' 'here'
Stop Word Removal	'guaranteed' 'return' '70%' 'investment' '$300' 'million' 'increase' 'economic' 'growth' 'year' 'After' 'breakup' 'largest' 'monopoly' 'America' 'earn' 'some' 'profit' 'Over' '70%' 'returns' 'annually' 'complete' 'details' 'just' 'click' 'here'
Stemming	'guarantee' 'return' '70%' 'investment' '$300' 'million' 'increase' 'economic' 'growth' 'year' 'After' 'breakup' 'largest' 'monopoly' 'America' 'earn' 'some' 'profit' 'Over' '70%' 'returns' 'annual' 'complete' 'details' 'just' 'click' 'here'
Classify Algorithm Output (NB or J48)	Email is Spam mail

5. Experimental results

In this section, the experimental results of the Spam Mail Detection (SMD) system are presented. An email dataset of total number of 1000 emails, 500 each of both the classifying algorithms are considered for the experimentation. A total number of three experiments are performed and evaluated results are compared. Two experiments are conducted for spam email detection using individual Naïve Bayes classification algorithm, J48 decision tree algorithm and third experiment is conducted by using hybrid bagged approach. Naïve Bayes algorithm is simple supervised learning algorithm which is easy to understand and implement. The algorithm shows good results even with small amount of training data. But the algorithm works with an assumption of dataset with independent class features. On the other hand, J48 is a decision tree based algorithm with the ability to handle feature interactions, missing values etc. But decision tree has difficulty in handling continuous data values along with the over-fitting issue. The hybrid bagged approach of Naïve Bayes and J48 algorithm assembled the best of both the algorithms. The overall result of spam mail detection system is the aggregation of the predictions of both the algorithms and thus ensuring a system with accurate and reliable results.

The efficiency of proposed spam mail detection system is encountered by evaluating the performance parameters. Parameters like precision, recall, accuracy, F-measure, true negative rate, false negative rate and false positive rate are calculated in order to evaluate the performance of the Spam mail Detection system. The performance evaluation of the SMD system is based on different measures presented in table 4.

The system has achieved an overall accuracy of 87.5% which is the average of the accuracies achieved by the two classifying algorithms. The accuracy achieved by Naïve Bayes classifier is 83.5% with precision and recall value of 85.26% and 81% respectively. Whereas, the accuracy achieved by J48 algorithm is 91.5% which with the precision and recall value of 93.68% and 89% respectively. Table 5 presents the evaluated results of the three experiments: Naïve Bayes, J48 algorithm and hybrid bagged approach respectively.

Table 4. Performance evaluation measures of SMD system

Evaluation Measure	Description	Formula
True Positive (TP)	No. of ham mails correctly identified.	N/A
False Positive (FP)	No. of spam mails incorrectly identified as ham.	N/A
True Negative(TN)	No. of spam mails correctly identified.	N/A
False Negative (FN)	No. of ham mails incorrectly identified as spam.	N/A
Precision	It defines the effectiveness of classifier.	$\frac{TP}{TP+FP}$
Recall (True Positive Rate)	Out of total class data, the positive labelled data returned by classifier.	$\frac{TP}{TP+FN}$
Accuracy	Ratio of the positive predicted values to the total data.	$\frac{TP+TN}{TP+TN+FP+FN}$
F-Measure	Overall performance by showing effective positive results by classifier.	$2\,.\,\frac{Precision\,.Recall}{Precision+Recall}$
True Negative Rate (TNR)	Ratio of correctly identified spam mails to total spam mails.	$\frac{TN}{TN+FP}$
False Negative Rate (FNR)	It identifies no of miss of spam mails.	$\frac{FN}{FN+TP}$
False Positive Rate (FPR)	Ratio of no. of spam mails incorrectly identified to the total no. of spam mails.	$\frac{FP}{FP+TN}$

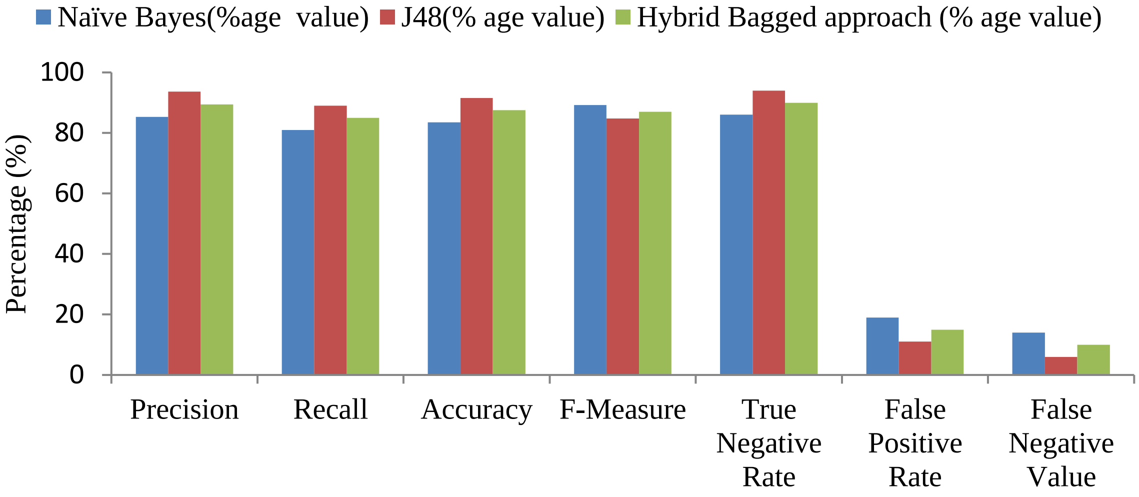

Figure. 5 Comprehensive results of SMD system, naïve bayes and J48

Table 5. Experimental results

Evaluation Measures	**Naïve Bayes**	**J48**	**Hybrid Bagged approach**
TP	81	89	85
FP	14	6	10
TN	86	94	90
FN	19	11	15
Precision (%)	85.26	93.68	89.47
Recall (%)	81	89	85
Accuracy (%)	83.5	91.5	87.5
F-Measure (%)	89.27	84.8	87.03
TNR (%)	86	94	90
FPR (%)	19	11	15
FNR (%)	14	6	10

The comparative analysis of the results as presented in table 5 clearly indicates that better results are achieved in terms of precision, recall and accuracy with J48 decision tree algorithm when compared with the Naïve Bayes and the hybrid bagged approach. However, the percentage value of F-measure in case of Naïve Bayes (89.27%) is higher than both the F-measure values of J48 (84.8%) and hybrid bagged approach (87.03%). Fig. 5 gives the graphical representation of the comparison of the result achieved by the SMD system and the corresponding classifying algorithms individually.

6. Conclusion

Spam email is one of the most demanding and troublesome internet issues in today's world of communication and technology. Spammers by generating spam mails are misusing this communication facility and thus affecting organisations and many email users. In this paper, a Spam Mail Detection system is introduced which makes use of a hybrid bagged approach for its implementation. The classification algorithms used in this approach are Naïve Bayes and J48. The accuracy achieved by Naïve Bayes and J48 algorithm is 83.5% and 91.5% respectively. The overall accuracy of 87.5% achieved by the hybrid bagged approach based SMD system shows that the experimental results are better when performed on only J48 algorithm. In order to enhance the system's performance and results, the concept of boosting approach could be considered for future work. The boosting technique will replace the weak classifier's learning features with the strong classifier's features and thus enhancing the overall system's performance.

References

[1] C. Yang, R. Harkreader, and G. Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers", In: *Recent Advances in Intrusion Detection*, Springer Berlin/Heidelberg, pp.318-337, 2011.

[2] S. Kumar, and S. Arumugam, "A Probabilistic Neural Network Based Classification of Spam Mails Using Particle Swarm Optimization Feature Selection", *Middle-East Journal of Scientific Research*, Vol.23, No.5, pp.874-879, 2015.

[3] N. P. DíAz, D. R. OrdáS, F. F. Riverola, and J. R. MéNdez, "SDAI: An integral evaluation methodology for content-based spam filtering models", *Expert Systems with Applications*, Vol.39, No.16, pp.12487-12500, 2012.

[4] A. K. Sharma, S. K. Prajapat, and M. Aslam, "A Comparative Study Between Naive Bayes and Neural Network (MLP) Classifier for Spam Email Detection", In: *IJCA Proceedings on National Seminar on Recent Advances in Wireless Networks and Communications. Foundation of Computer Science (FCS)*, pp.12-16, 2014.

[5] W. Ma, D. Tran, and D. Sharma, "A novel spam email detection system based on negative selection", In: *Proc. of Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT'09*, Seoul, Korea, pp.987-992, 2009.

[6] T. S. Guzella, and W. M. Caminhas, "A review of machine learning approaches to spam filtering", *Expert Systems with Applications*, Vol.36, No.7, pp.10206-10222, 2009.

[7] G. Chandrashekar, and F. Sahin, "A survey on feature selection methods", *Computers & Electrical Engineering*, Vol.40, No.1, pp.16-28, 2014.

[8] M. Mohamad, and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification", In: *Proc. of 2015 International Conference on Computer, Communications, and Control Technology (I4CT)*, Kuching, Sarawak, Malaysia, pp.227-231, 2015.

[9] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm", In: *Proc. of 2014 International Conference on Optimization, Reliabilty, and Information Technology (ICROIT)*, Faridabad, Haryana, pp.153-155, India, 2014.

[10] S. Youn, and D. McLeod, "Efficient spam email filtering using adaptive ontology." In: *Proc. of Fourth International Conference on Information Technology*, Las Vegas, NV, USA, pp.249-254, 2007.

[11] H. Faris, and I. Aljarah, "Optimizing feedforward neural networks using Krill Herd algorithm for e-mail spam detection", In: *Proc. of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT),* Amman, Jordan, pp.1-5, 2015.

[12] H. Faris, I. Aljarah, and B. Al-Shboul, "A Hybrid Approach Based on Particle Swarm Optimization and Random Forests for E-Mail Spam Filtering", In: *Proc. of International Conference on Computational Collective Intelligence*, Halkidiki, Greece, pp.498-508, 2016.

[13] S. K. Tuteja, "A Survey on Classification Algorithms for Email Spam Filtering", *International Journal of Engineering Science,* Vol.6, No.5, pp.5937-5940, 2016.

[14] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues", *IEEE Access*, Vol. 5, pp. 9044-9064, 2017.

[15] S. Ajaz, M. T. Nafis, and V. Sharma, "Spam Mail Detection Using Hybrid Secure Hash Based Naive Classifier", *International Journal of Advanced Research in Computer Science*, Vol.8, No.5, pp.1195-1199, 2017.

[16] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive Bayes for text categorization", *IEEE Transactions on Knowledge and Data Engineering*, Vol.28, No.9, pp.2508-2521, 2016.

[17] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited", In: *Proc. of Australasian Joint Conference on Artificial Intelligence*, Vol.3339, Cairns, Australia, pp. 488-499, 2004.

[18] S. Youn, and D. McLeod, "A comparative study for email classification", *Advances and innovations in systems, computing sciences and software engineering*, pp.387-391, Springer, Dordrecht, 2007.

[19] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering", In: *Proc. of the workshop on Machine Learning in the New Information Age*, Barcelona, Spain, pp.9-17, 2000.