Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018)
IEEE Xplore Compliant Part Number: CFP18K74-ART; ISBN:978-1-5386-2842-3

Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization

Kriti Agarwal
Computer Science and Engineering
Radha Govind Engineering College, Meerut, India
kritiag94@gmail.com

Tarun Kumar
Computer Science and Engineering
Radha Govind Engineering College, Meerut, India
tarunrggi@gmail.com

Abstract- Now-a-days, communication through email has become one of the cheapest and easy ways for the official and business users due to easy availability of internet access. Most of the people prefer to use email to share important information and to maintain their official records. But just like the two sides of coin, many people misuse this easy way of communication by sending unwanted & useless bulk emails to others. These unwanted emails are spam emails that affect the normal user to face the problems like excessive usage of their mailbox memory and filtration of useful email from unwanted useless emails. So, there is the need of some autonomous approach that filters the excessive data of emails in the form of spam emails. In this paper, an integrated approach of machine learning based Naïve Bayes (NB) algorithm and computational intelligence based Particle Swarm Optimization (PSO) is used for the email spam detection. Here, Naïve Bayes algorithm is used for the learning and classification of email content as spam and non-spam. PSO has the stochastic distribution & swarm behavior property and considered for the global optimization of the parameters of NB approach. For experimentation, dataset of Ling spam dataset is considered and evaluated the performance in terms of precision, recall, f-measure and accuracy. Based on the evaluated results, PSO outperforms in comparison with individual NB approach.

***Keywords**- Email Spam Detection, Text Mining, Machine Learning, Naïve Bayes, Swarm Intelligence, Particle Swarm Optimization.*

I. INTRODUCTION

Electronic Mails (emails) are the easy and efficient way adapted from individual user to business organization to share knowledgeable and important information [1]. But many people spamming its content by sending phishing, spoofing & useless junk emails. It is difficult for a user to filter out useful emails manually from daily receiving of 40-50 emails which contains almost 60-70% email spam [2]. The email spam is increasing upto the extent that many users send spam content to hack someone's account details, to create more network traffic, and to waste someone's energy & time. Technical hackers have availability of various code syntaxes (like HTML tree code) to block or hack someone's email account or to modify the email contents. Researchers are continuously working on the methods and algorithms to filter out these email spam. Email spam filtration is a text mining based approach to classify the available email text content into spam and non-spam emails. Different authors have used different machine learning based and computational intelligence based approaches for the detection and classification of email spam with experimentation on different datasets. The available methods and techniques used by different researchers are discussed in section 2.

In this paper, an integrated approach of machine learning based Naïve Bayes (NB) approach and computational intelligence based Particle Swarm Optimization (PSO) is considered for the email spam detection. NB is based on the Bayes theorem having strong independence and probability distribution property. PSO is swarm intelligence concept inspired from the social behavior of flying birds and fishes etc. For experimentation, 1000 emails from the Ling spam dataset [3] are considered. Experimental results are evaluated for NB and integrated approach of NB & PSO in terms of precision, recall, f-measure and accuracy.

Rest of the paper is organized in the following manner: Section 2 presents the existing work related to email spam classification. Section 3 presents the basic concepts of Naïve Bayes and Particle Swarm Optimization. Section 4 presents the discussion on the proposed integrated approach of NB and PSO along with the working flow chart of proposed concept. Section 5 presents the used dataset, evaluation parameters, and evaluated results for NB & proposed approach along with the comparison. Section 6 concludes the paper along with some future directions.

II. RELATED WORK

In this section, the existing techniques and methods used by researchers for email spam classification are presented. Different methods have been adapted by different authors with experimentation on different datasets. Table I illustrates the work of different authors along with the information of used methods, datasets, and remarks.

Some authors have worked on the detection of spam email in both the textual and image data format. Harisinghaney et al. (2014) [4] and Mohamad & Selamat (2015) [5] have used the image and textual dataset for the email spam detection with the use of different methods. Harisinghaney et al. (2014) have used methods of Naïve Bayes, KNN algorithm and Reverse DBSCAN algorithm with experimentation on Enron Corpus's dataset. For the text recognition, OCR library is used but this OCR does not perform well. Two experiments are performed with and without preprocessing steps with evaluation parameters of accuracy, specificity, sensitivity and precision. Overall Naïve Bayes perform efficiently with accuracy of 87%. Further, Mohamad & Selamat (2015) considered feature selection hybrid approach of TF-IDF (Term Frequency Inverse Document Frequency) and Rough set theory. In this experimentation, a manual dataset of 169 emails is generated containing both the text and images based data. Authors have used the RSES (Rough Set Exploration System) tool for the removal of unessential words and for the rules generation. Overall concept is compared in terms of accuracy to TDIDF-Decision Tree and shows efficient results.

Most of the researchers have focused only on the text based email spam classification as image based spam can be filtered at the initial stage of pre-processing. For textual data mining, major use of machine learning based Support Vector Machine (SVM) is recorded. Either some of the authors have used SVM individually (Renuka and Visalakshi, 2014) [6] or some have used SVM in integration with some other concepts like SVM-NB (Feng et al., 2016) [7], SCS-SVM (Kumaresan and Palanisamy, 2017) [8], and SVM-ELM (Olatunji et al., 2017) [9]. In 2014, Renuka and Visalakshi have used Support vector Machine (SVM) for the classification of Email Spam detection along with the use of Latent Semantic Indexing (LSI) for feature selection. TF-IDF is used for the feature extraction. Here, proposed SVM-LSI is compared with SVM-TFIDF without using LSI, PSO and Neural Network. From the considered methods, SVM-LSI performs better in terms of accuracy as compare to other existing concepts. In 2016, Feng et al. have proposed SVM-NB algorithm for the email spam filtration. Authors have combined the SVM algorithm with NB approach where NB can handle large dataset and SVM is able to create hyper-plane based separation between different feature categories. Authors have used Chinese Spam Email Dataset obtained from DATAMALL. From evaluated results, authors reported the better performance of proposed SVM-NB approach as compare to individual SVM and NB. In 2017, Kumaresan and Palanisamy have modified the concept of Cuckoo Search with stepsize and Support Vector Machine (SVM) for the Email Spam Classification. Modified Cuckoo Search approach is named as Stepsize Cuckoo Search (SCS) by the authors. In this email spam classification, SCS approach is used to select the best feature set and SVM is used to calculate fitness value & final classification. Overall performance is evaluated for three kernels of Linear, Quadratic and Polynomial of SVM with evaluation parameters of Specificity, Sensitivity and Accuracy in comparison to original CS along with SVM. Evaluated results shows that proposed SCS with SVM is better as compare to CS with SVM. Further, in the same year 2017, Olatunji et al. used SVM with Extreme Learning Machines (ELM) for the Email Spam Classification. ELM is machine learning approach that was proposed to overcome the perennial drawback of feed forward neural network and is used as learning approach for single layer based neural network. Results of ELM and SVM are compared on the basis of Accuracy and Time taken for the email spam classification from same dataset. In terms of Accuracy, SVM performs better with 94.06 % as compare to ELM having accuracy 93.04 %. But for each case, SVM consumes more time as compare to ELM. So, ELM is better than SVM in terms of time taken.

Apart from the consideration of SVM approach, there is the consideration of other papers in which methods of PSO and some other techniques are also used. In 2015, Idris et al. [10] worked on the improvement of Negative Selection Algorithm (NSA) with Particle Swarm Optimization (PSO) for the email spam classification. Here, random detector of NSA approach is improved with PSO approach. Overall fitness function of NSA-PSO is evaluated with Local Outlier Factor (LOF). Overall NSA-PSO achieved 83.20 % accuracy which is much higher as compare to accuracy of 68.86%. In 2016, Tuteja and Bogiri [11] have used Artificial Neural Network based concept for the classification and identification of spam emails from manually created dataset. In this process K-means clustering is used for the feature extraction, Back Propagation Neural Network is used for the training of dataset and Feed Forward Neural Network for the classification & identification of email spam. Results with k-means clustering in preprocessing are better as compare to without using k-means clustering approach. Further, in 2016, Kaur and Sharma [12] have integrated the concept of Particle Swarm Optimization (PSO) with Decision Tree Algorithm to improve the Email Spam Classification. Here, PSO is considered to evaluate the end results of any specific dataset and Decision Tree is used to classify complete dataset into small feature classes. Overall efficient results are obtained as compare to other concept. The major drawback of the concept is that authors have not used any feature extraction technique.

Table I
Existing work related to email spam detection

Author and Year	Dataset Used	Method Used	Evaluation Parameters	Remarks
Renuka and Visalakshi (2014)	Ling Spam Email Corpus	Support vector Machine (SVM) with Latent Semantic Indexing (LSI)	Precision, recall and accuracy	Proposed SVM-LSI is performs well in comparison with other state of art research.
Harisinghaney et al. (2014)	Enron Corpus dataset	Naïve Bayes, KNN algorithm and Reverse DBSCAN algorithm	Precision, specificity, sensitivity and accuracy	Reported better results with the pre-processing steps in comparison with results without pre-processing steps. But huge time consumption is recorded by authors for data

				filtration.
Idris et al. (2015)	Spam base dataset	Negative Selection Algorithm and Particle Swarm Optimization	Negative prediction value, accuracy, F-measure and statistical t-test, specificity, sensitivity, correlation factor and positive prediction value	Overall NSA-PSO achieved better accuracy results as compare to accuracy NSA
Mohamad and Selamat (2015)	Manually Generated Dataset	Term Frequency Inverse Document Frequency and Rough set theory	Classification accuracy	Major focus was on feature extraction approach instead of email spam classification
Tuteja and Bogiri (2016)	Manually Generated Dataset	K-means Clustering and Artificial Neutral Network	Precision and recall	Observed better results with preprocessing steps in comparison with results with without pre-processing
Kaur and Sharma (2016)	Spam base dataset	Integrated Concept of PSO and Decision Tree Algorithm	F-measure, mean absolute error and correctly classified ratio	There is no information about the use of feature extraction approach.
Feng et al. (2016)	DATA MALL (Chinese Spam Email Dataset)	Integrated SVM-NB (Support Vector Machine -Naïve Bayes)	Precision, recall and execution time	Integrated approach improves the results in comparison with individual SVM and NB approaches
Kumaresan & Palanisamy (2017)	Ling-Spam dataset	Stepsize Cuckoo Search with Support Vector Machine	Accuracy, specificity, and sensitivity	Overall modified algorithm is superior and speeds up the classification as compare to original CS
Olatunji et al. (2017)	Text Corpus Spambase dataset	Extreme Learning Machines and Support Vector Machines	Accuracy and time taken	SVM performs better than ELM in terms of accuracy but ELM takes less time than SVM

III. BASIC CONCEPTS

In this section, the basic of Naïve Bayes algorithm and Particle Swarm Optimization is discussed as there concepts are used as an integrated approach for the email spam classification.

A. Naïve Bayes Approach

Naïve Bayes algorithm [13] [14] is a Bayes theorem based statistical machine learning based approach having properties of strong independence, probability distribution and ability to handle large datasets. In NB, probability distribution is evaluated from the frequency distribution of dataset. In classification problem, bayes decision rule is used to assign a class. As per bayes' decision rule, class having highest value of posterior probability is chosen by the classifier. The posterior probability can be evaluated with Equation (1).

$$P(y|x) = \frac{P(x|y)\ P(y)}{P(x)} \qquad \text{…Equation (1)}$$

Where, x is any feature vector set $(x_1, x_2, x_3, ...x_n)$ and y are the class variables with m possible outcomes $(y_1, y_2, y_3, ...y_n)$. $P(y|x)$ stands for posterior probability, $P(x|y)$ is any particular class on which $P(y|x)$ is dependent. $P(x)$ is evidence depending on the known feature variables, $P(y)$ is the prior probability. So, Naïve Bayes classification model consists of set of probabilities of prior probability, class conditional probability and posterior probability.

B. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is swarm intelligence based concept derived in 1995 by Eberhart and Kennedy [15] by getting the inspiration from the social behavior of flying birds and school of fishes. PSO work on the property of stochastic distribution and initially find the local search solution, then individual particle share their solution and global solution is obtained. This property is known as global optimization property. This algorithm works in an iteration manner and moves closer to the best solution. Initially, particles begin the process by the casual fly in the form of population of N particle solution. In the S-dimensional space, the position of the i^{th} particle is represented as a point in this space, where S is the number of variables participated. In the entire process, particles try to find the global best solution.

PSO algorithm works on the basis of two main dynamic vectors particle position and velocity that changes according to interactions between the different particles as each particle represent a solution. Each particle have ability to change their trajectory as per the experience and sharing properties with other particles to achieve better solution with each increase in iteration. In last few decades, PSO has been used by different authors to solve different problems in different domains [16] [17] [18]. In this research article, PSO is considered to optimize the classification results attained using NB approach for email spam detection.

IV. Proposed Algorithm

In this section, an integrated concept of NB and PSO is presented. NB having probability distribution property determines the possible class for the email content from the spam class or non-spam class on the basis of keywords present in the email textual data. PSO is used to further optimize the parameters of NB approach to improve the accuracy, search space and classification process. Here, correlation based feature selection (CFS) is used to select the relevance features from the bags of words on which basis classification is performed. The work flow of proposed concept is shown figure 1. To better understand the process, the stepwise algorithm is presented for an individual email below:

Step 1: Consider a random email from the ling spam dataset for experimentation.

Step 2: The considered email is in raw form. To perform the feature extraction/selection and classification procedure, initially email is needed to pre-process. Pre-processing involves the steps of tokenization, stemming and stop word removal.

2.1. Initially, tokenize the email into individual keywords. Tokenization split each individual word into different token.

2.2. Remove the stop words from the obtained tokens.

2.3. Perform stemming on the tokens obtained from the previous step. Stemming process reduces the size of word to its root word. For stemming, a predefined list of possible words with their respective stem words is considered.

2.3.1. For stemming, a list of suffix words is stored into array with their respective root words.

2.3.2. Consider check_token = availability in considered array of root word

2.3.3. If suffix of check_token = true, stem the word to its respective root word from the list of array.

2.3.4. Else, there is no need of stemming. Word is already in its root word format. Move to next token.

Step 3: Apply Correlation based feature selection approach to select the useful feature words from the pre-processed data. 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 CFS can be applied as mentioned in Equation 2.

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

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

Step 4: Calculate the probability distribution of the tokens along with selected features using NB approach. The formulation for the probability distribution is presented in Equation 3.

$$P(y|(f1, f2, f3, \dots . fn)) = \frac{P((f1, f2, f3, \dots . fn)|y)\ P(y)}{P((f1, f2, f3, \dots . fn))} \quad \text{...Equation (3)}$$

Where, *f* is any feature vector set ($f_1, f_2, f_3, \dots . f_n$) and *y* are the class variables with *m* possible outcomes ($y_1, y_2, y_3, \dots . y_n$). *P(y|x)* stands for posterior probability, *P(x|y)* is any particular class on which *P(y|x)* is dependent. *P(x)* is evidence depending on the known feature variables, *P(y)* is the prior probability.

Step 5: Apply PSO approach to optimize the parameters of NB approach.

5.1. All the tokens are considered as the particles. Initially these particles randomly fly and search for the food sources in the form of best feature match for tokens. Then search for the local and global solution.

5.2. The performance of each particle depends upon the similarity with the features that has to optimize.

5.3. Each particle flies over the n- dimensional outer search space and keep updating the following information:

- X_i – current position of particle
- P_i – the personal best position of particle
- V_i – the current velocity of particle

5.4. The velocity updates in PSO can be calculated using the formula given below by Equation (4):

$$V_{i(t+1)} = \omega\, V_{i(t)} + c_1 r_1 \left(P_{i(t)} - X_{i(t)}\right) + c_2 r_2 \left(P_g - X_{i(t)}\right) \quad \text{...Equation (4)}$$

Now, V_i is the new velocity. So, the position of the particle updates with the velocity as defined with Equation (5):

$$X_{i(t+1)} = X_{i(t)} + V_{i(t+1)} \quad \text{....Equation (5)}$$

5.5. Update the positions for each particle and store the global best solutions.

Step 6: Based on the evaluated feature similarity using PSO, classification of tokens is declared as spam or non spam.

Step 7: Further, final classification is performed by evaluating the probability of spam or non-spam tokens in sentence.

7.1. If probability value of spam tokens is more, then email is considered as spam email.

7.2. Else, email is considered as non-spam email.

Step 8: Store the email as spam or non-spam and repeat the process for all the emails.

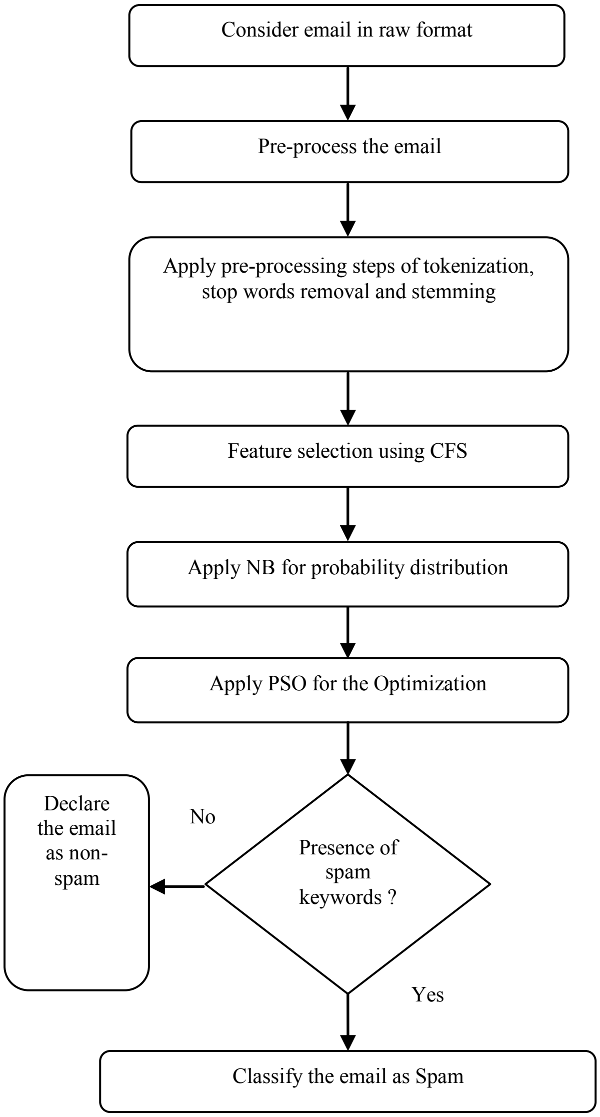
Consider email in raw format
Pre-process the email
Apply pre-processing steps of tokenization, stop words removal and stemming
Feature selection using CFS
Apply NB for probability distribution
Apply PSO for the Optimization
Presence of spam keywords ?
No
Declare the email as non-spam
Yes
Classify the email as Spam

Figure 1: Flowchart of email spam detection

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the used dataset, performance evaluation measures, evaluated results, and comparison with the individual NB approach.

A. Dataset Used

The proposed concept is experimented on the dataset of Ling Spam dataset [3]. From Ling spam dataset 1000 randomly selected emails are used. From these 1000 emails, 600 emails are used for training and 400 emails for testing by maintaining a ratio of 60:40. Out of 600 training emails, 300 emails are spam and 300 emails are non-spam. In the similar manner, out of 400 testing emails, 200 emails are spam and 200 are non-spam. Initially, training step is performed using NB and proposed integrated approach of NB & PSO. Then, based on the testing emails, results are evaluated for individual NB and proposed integrated approach of NB & PSO.

B. Evaluation Parameters

Performance of proposed algorithm is evaluated in terms of precision, recall, f-measure and classification accuracy. These parameters can be calculated with the help of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These measures are defines as below.

TP can be defined as the numbers of spam emails are correctly identified as spam.

TN can be defined as the numbers of non-spam emails are correctly identified as non-spam.

FP can be defined as the numbers of non-spam emails are incorrectly identified as spam.

FN can be defined as the numbers of spam emails are incorrectly identified as non-spam.

Precision can be defined as the probability of spam email detection with true value using the classifier. It also defines the effectiveness of classifier. It can be formulated as mentioned in Equation 6.

$$P = \frac{TP}{TP+FP} \qquad \ldots\text{Equation (6)}$$

Recall can be defines as the probability of actual detection of email spam. It can be formulated as mentioned in Equation 7.

$$R = \frac{TP}{TP+FN} \qquad \text{....Equation (7)}$$

F-Measure is a measure to define the overall performance of the classifier. It is evaluated from the precision and recall values as mentioned in Equation 8.

$$F = \frac{2PR}{P+R} \qquad \ldots\text{Equation (8)}$$

Accuracy can be defined as the ratio of positive predicted values to total data values. It can be evaluated as mentioned in Equation 9.

$$A = \frac{TP+TN}{TP+TN+FP+FN} \ldots \text{ Equation (9)}$$

C. Results and Comparison

Based the above mentioned formulations presented in Equation (6), Equation (7), Equation (8), and Equation (9), values of precision, recall, f-measure and accuracy are evaluated. Values of TP, TN, FP and FN are important as precision, recall, f-measure and accuracy are calculated based on the values of TP, TN, FP and FN. The calculated values of TP, TN, FP and FN using individual NB and integrated proposed concept are shown in table II. Further, calculated values of precision, recall, f-measure and accuracy for individual NB and integrated proposed concept are shown in table III and comparison graph in figure 2.

Table I : Evaluated value of TP, TN, FP, and FN

Evaluation Measures	Naïve Bayes	Proposed Integrated Concept
TP	173	189
FN	27	11
TN	178	193
FP	22	07

Table II: Evaluated value of TP, TN, FP, and FN

Evaluation Measures	Naïve Bayes	Proposed Integrated Concept
Precision (%)	88.71 %	96.42 %
Recall (%)	86.50 %	94.50 %
F-Measure (%)	87.59 %	95.45 %
Accuracy (%)	87.75 %	95.50 %

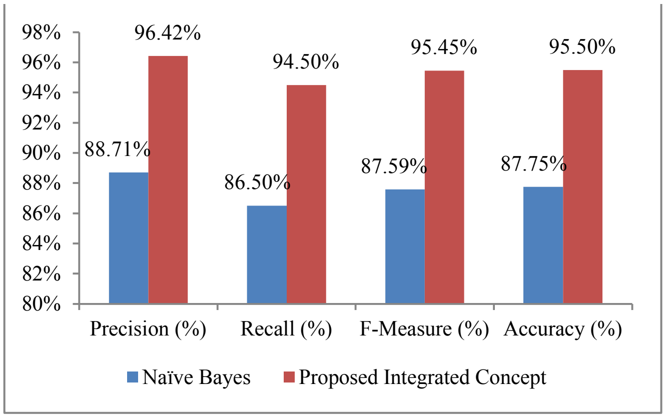

Figure 2: Comparison of proposed integrated concept with individual NB approach

From the comparison figure 2, it can be seen that proposed integrated approach of NB & PSO outperforms in comparison with individual NB approach. The classification accuracy for the individual NB approach lacks with 7.75 % from proposed integrated concept of NB & PSO. The main advantage of integrated concept is the availability of optimization technique of PSO that have the ability to optimize the solution with global search solution space.

VI. CONCLUSION

Email spam has become one of the most demanding research topics due to increasing cyber crime and increasing spammers. Different authors have used different methods with testing on different datasets to detect the email spam as discussed in section 2. With analysis from the results of existing techniques, we have used integrated approach of NB and PSO for the email spam detection. NB having probability distribution property determines the possible class for the email content from the spam class or non-spam class on the basis of keywords present in the email textual data. PSO is used to further optimize the parameters of NB approach to improve the accuracy, search space and classification process. Correlation based feature selection (CFS) is used as a feature selection approach. Experimentation is performed on the Ling spam dataset with the use of evaluation parameters of precision, recall, f-measure and accuracy. From the evaluated results, it can be declared that proposed integrated concept outperformed in comparison with individual NB approach. For future directions, Naïve Bayes approach can also be integrated with any other swarm optimization based concept like ant colony optimization, artificial bee colony optimization, firefly algorithm etc. Also, NB approach can also be changed with any other machine learning based algorithm to further improve the results.

REFERENCES

Lucas, William. "Effects of e-mail on the organization." *European Management Journal* 16, no. 1 (1998): 18-30.

[1] Team, Radicati. "Email Statistics Report, 2015-2019. The Radicati Group." (2015).

[2] Androutsopoulos I., J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering", In: *11th European Conference on Machine Learning*, pp.9-17, Barcelona, Spain, 2000.

[3] Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In *Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on*, pp. 153-155. IEEE, 2014

[4] Mohamad, Masurah, and Ali Selamat. "An evaluation on the efficiency of hybrid feature selection in spam email classification." In *Computer, Communications, and Control Technology (I4CT), 2015 International Conference on*, pp. 227-231. IEEE, 2015.

[5] Renuka, Karthika D., and P. Visalakshi. "Latent Semantic Indexing Based SVM Model for Email Spam Classification." (2014).

[6] 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.

[7] Kumaresan, T., and C. Palanisamy. "E-mail spam classification using S-cuckoo search and support vector machine." *International Journal of Bio-Inspired Computation* 9, no. 3 (2017): 142-156.

[8] Olatunji, Sunday Olusanya. "Extreme Learning machines and Support Vector Machines models for email spam detection." In *Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on*, pp. 1-6. IEEE, 2017.

[9] Idris, Ismaila, Ali Selamat, Ngoc Thanh Nguyen, Sigeru Omatu, Ondrej Krejcar, Kamil Kuca, and Marek Penhaker. "A combined negative selection algorithm–particle swarm optimization for an email spam detection system." *Engineering Applications of Artificial Intelligence* 39 (2015): 33-44.

[10] 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.

[11] Kaur, Harpreet, and Ajay Sharma. "Improved email spam classification method using integrated particle swarm optimization and decision tree." In *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*, pp. 516-521. IEEE, 2016.

[12] Murphy, Kevin P. "Naive bayes classifiers." *University of British Columbia* 18 (2006).

[13] Cichosz, Paweł. "Naïve Bayes classifier." *Data Mining Algorithms: Explained Using R* (2015): 118-133.

[14] Eberhart, Russell, and James Kennedy. "A new optimizer using particle swarm theory." In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, pp. 39-43. IEEE, 1995.

[15] Shi, Yuhui. "Particle swarm optimization: developments, applications and resources." In *evolutionary computation, 2001. Proceedings of the 2001 Congress on*, vol. 1, pp. 81-86. IEEE, 2001.

[16] Rana, Sandeep, Sanjay Jasola, and Rajesh Kumar. "A review on particle swarm optimization algorithms and their applications to data clustering." *Artificial Intelligence Review*35, no. 3 (2011): 211-222.

[17] Parsopoulos, Konstantinos E., ed. *Particle swarm optimization and intelligence: advances and applications: advances and applications*. IGI global, 2010.